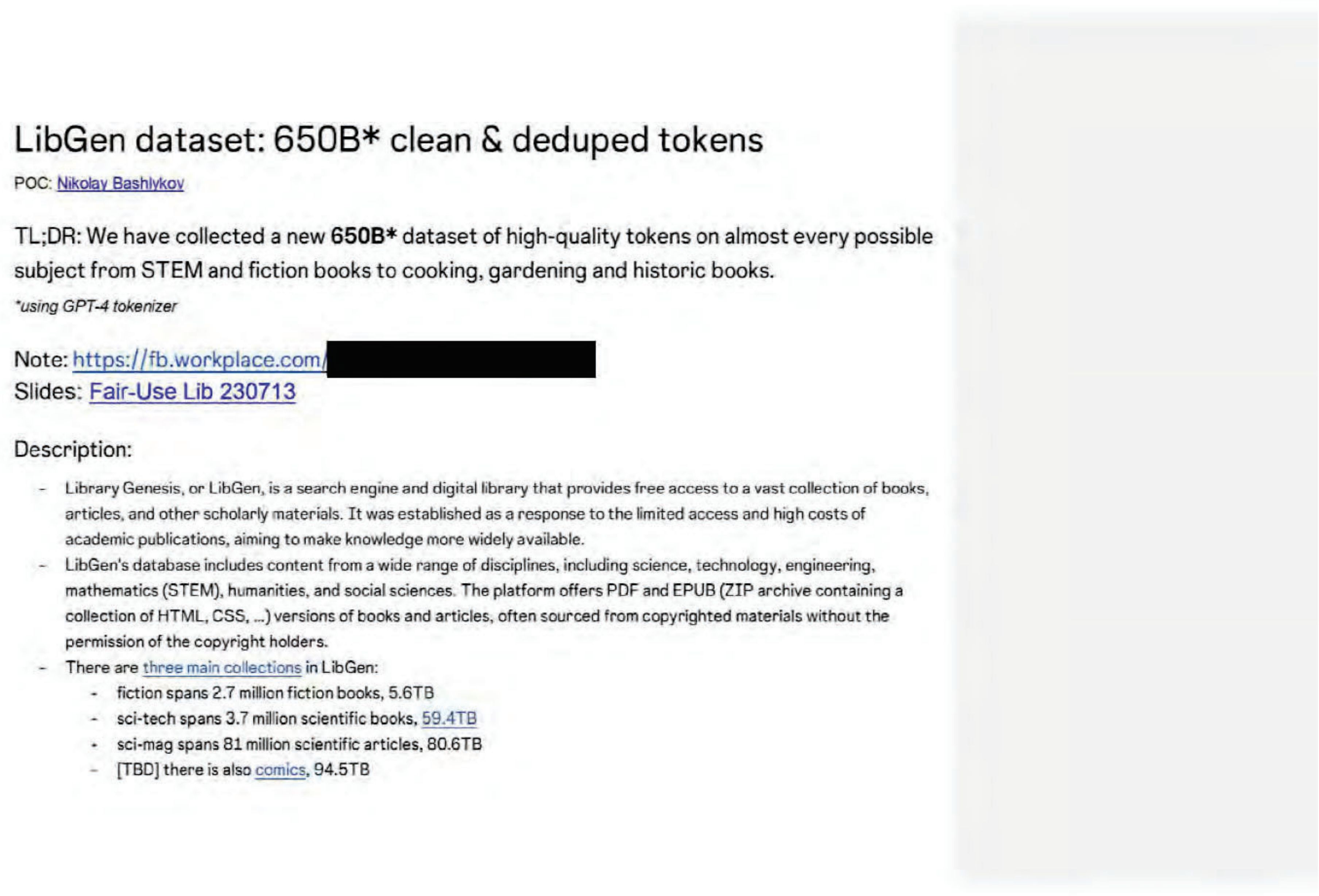
EXHIBIT D



LibGen dataset: 650B* clean & deduped tokens

POC: Nikolay Bashlykov

TL;DR: We have collected a new **650B*** dataset of high-quality tokens on almost every possible subject from STEM and fiction books to cooking, gardening and historic books.

**using GPT-4 tokenizer*

Note: https://fb.workplace.com/[redacted]
Slides: Fair-Use Lib 230713

Description:

- Library Genesis, or LibGen, is a search engine and digital library that provides free access to a vast collection of books, articles, and other scholarly materials. It was established as a response to the limited access and high costs of academic publications, aiming to make knowledge more widely available.
- LibGen's database includes content from a wide range of disciplines, including science, technology, engineering, mathematics (STEM), humanities, and social sciences. The platform offers PDF and EPUB (ZIP archive containing a collection of HTML, CSS, ...) versions of books and articles, often sourced from copyrighted materials without the permission of the copyright holders.
- There are three main collections in LibGen:
  - fiction spans 2.7 million fiction books, 5.6TB
  - sci-tech spans 3.7 million scientific books, 59.4TB
  - sci-mag spans 81 million scientific articles, 80.6TB
  - [TBD] there is also comics, 94.5TB

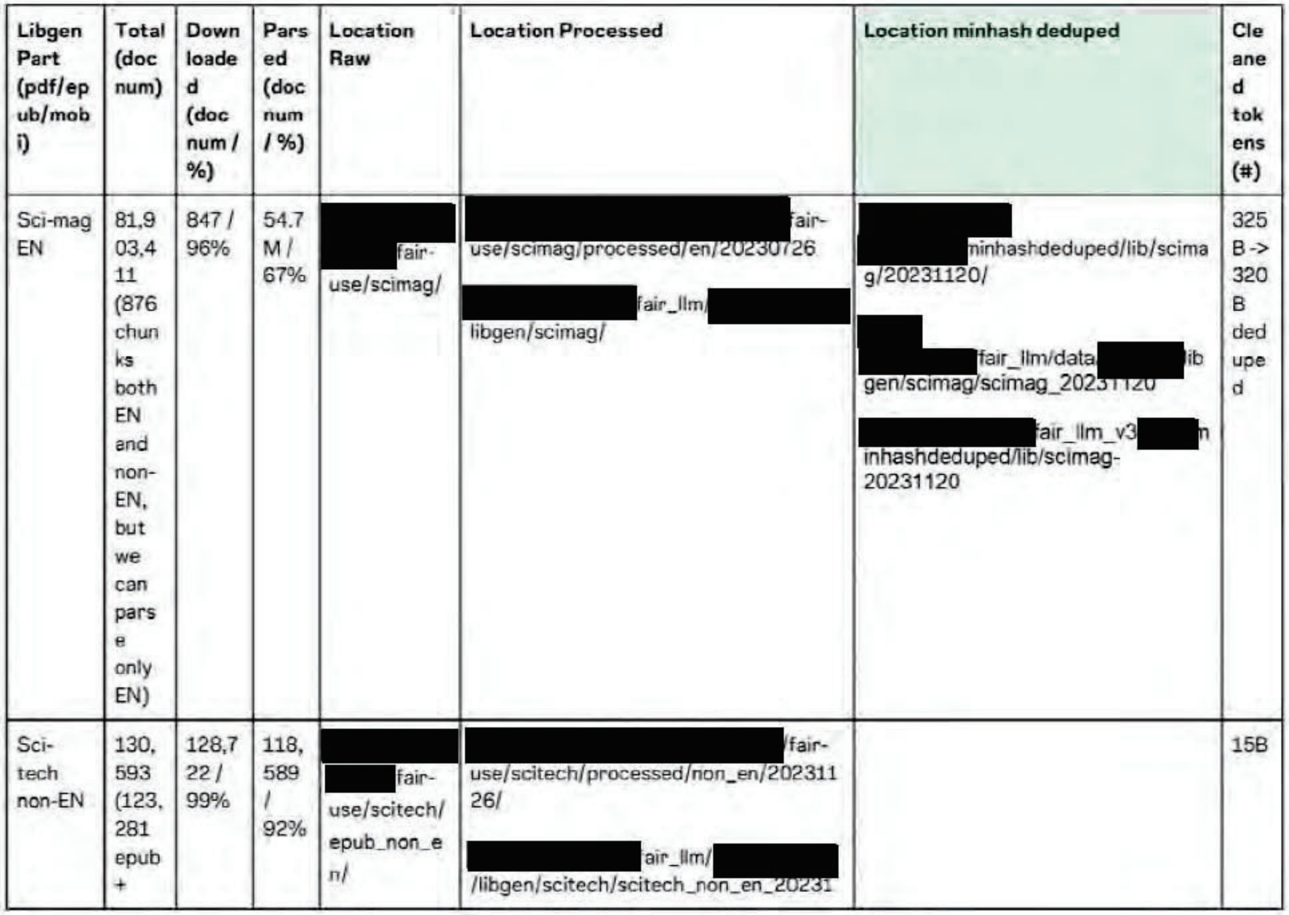
- Analogues:
    - Sci-Hub: similar to the sci-mag part of LibGen.
    - Z-lib: initially a mirror of LibGen, but then evolved to a separate project. Now claims to have 23M books and 285B articles. Banned multiple times, but seems to be working currently. Worth investigating.

The PDFs are parsed with the NOUGAT library

LibGen (full DB)	fiction	sci-tech	sci-mag	Total
Total documents (#)	2,693,056	3,706,772	81,903,411	
Unique documents (author&title)	1,607,593	3,274,071	72,624,976	
Language (%)	English: 65% German: 11% French: 6%	English: 51% Russian 29% German: 5%	N/A	
Format (%)	Epub: 59% PDF: 11% mobi: 10%	Epub: 16% PDF: 65% djvu: 11%	PDF: ~100%	
Median number of pages per doc (#)	170	258	6	
Extracted EN clean tokens (#)	110B	220B	325B	
Deduped EN tokens (gpt-4 tokenizer)	70B	190B	320B	
Extracted non-EN clean tokens (#)	55B	15B	-	
Extracted ALL clean&deduped tokens (#)	**125B**	**205B**	**320B**	**650B**

SUMMARY TABLE

Libgen Part (pdf/epub/mobi)	**Total (doc num)**	**Downloaded (doc num / %)**	**Parsed (doc num / %)**	**Location Raw**	**Location Processed**	**Location minhash deduped**	**Cleaned tokens (#)**
Sci-tech EN	1,726,719 (454,064 epubs + 1,272,655 pdfs)	1,695,684 / 98%	1,496,473 / 88%	[redacted]/fair-use/scitech/	[redacted]/fair-use/scitech/processed/en/20230526/ [redacted]fair_llm/[redacted]libgen/scitech/scitech_en	[redacted]minhashdeduped/lib/scitech/20231120/ [redacted]fair_llm/[redacted] [redacted]libgen/scitech/scitech_en_20231120 [redacted]fair_llm_v3[redacted]m inhashdeduped/lib/scitech-20231120	220B -> 190B deduped
Fiction EN	1,159,720 (1,041,740 epubs + 117,980 pdfs)	1,138,296 / 98%	1,042,125 / 92%	[redacted]fair-use/fiction/	[redacted]/fair-use/fiction/processed/en/20230526/ [redacted]fair_llm[redacted]shuffled/libgen/fiction/fiction_en	**Fiction safe (w/o adult content)** [redacted]minhashdeduped/lib/fiction/20231210/safe **Fiction rest (w adult content)** [redacted]minhashdeduped/lib/fiction/20231210/rest [redacted]fair_llm_v3[redacted]m inhashdeduped/lib/fiction-20231210	110B -> 70B deduped



Libgen Part (pdf/epub/mobi)	Total (doc num)	Downloaded (doc num / %)	Parsed (doc num / %)	Location Raw	Location Processed	Location minhash deduped	Cleaned tokens (#)
Sci-mag EN	81,903,411 (876 chunks both EN and non-EN, but we can parse only EN)	847 / 96%	54.7M / 67%	[redacted] fair-use/scimag/	[redacted] fair-use/scimag/processed/en/20230726 [redacted] fair_llm/[redacted] libgen/scimag/	[redacted] minhashdeduped/lib/scimag/20231120/ [redacted] fair_llm/data[redacted] lib gen/scimag/scimag_20231120 [redacted] fair_llm_v3[redacted] minhashdeduped/lib/scimag-20231120	325B -> 320B deduped
Sci-tech non-EN	130,593 (123,281 epub +	128,722 / 99%	118,589 / 92%	[redacted] fair-use/scitech/epub_non_en/	[redacted] fair-use/scitech/processed/non_en/20231126/ [redacted] fair_llm/[redacted] /libgen/scitech/scitech_non_en_20231		15B

Libgen Part (pdf/epub/mobi)	Total (doc num)	Downloaded (doc num / %)	Parsed (doc num / %)	Location Raw	Location Processed	Location minhash deduped	Cleaned tokens (#)
	7,312 mobi)				126 [redacted]fair_llm_v3/[redacted]_v3/fair-use/scitech/processed/non_en/scitech-20231126		
Fiction non-EN	594,348 (545,578 epub + 48,770 mobi)	586,240 / 99%	461,246 / 79%	[redacted]fair-use/fiction/epub_non_en	[redacted]fair-use/fiction/processed/non_en/20231126/ [redacted]/fair_llm/[redacted]/libgen/fiction/fiction_non_en_20231126 [redacted]air_llm_v3/data/data_v3/fair-use/fiction/processed/non_en/fiction-20231126		55B
Total (gpt-4 tokenizer)							**725 B -> 650 B ded**

Libgen Part (pdf/epub/mobi)	Total (doc num)	Downloaded (doc num / %)	Parsed (doc num / %)	Location Raw	Location Processed	Location minhash deduped	Cleaned tokens (#)
							uped

Updates:

26.11.2023

Multilingual LibGen v2

Similar cleaning steps were applied to multilingual libgen (fiction and scitech) as well, except for token distribution KL divergence heuristics.

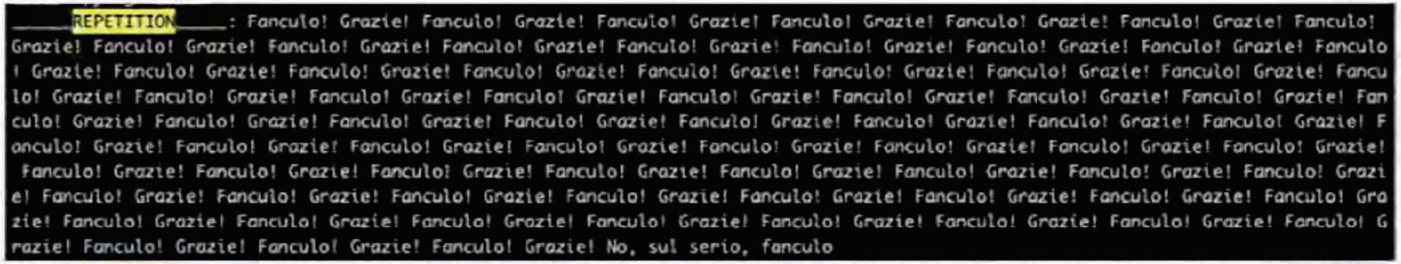
- We did not apply the token distribution outliers heuristics because the top documents returned by high KL divergence do not show clear patterns of repetition or ungrammatical text in multilingual libgen. Part of the reason is that we concatenated all non-English documents together, so the corpus is not homogenous for the tool to be useful. We decided to skip this step for multilingual in the short term, and we can revisit it later when we split the data by language.

- Overall, we removed **1%** and **0.67%** of total characters from fiction and scitech respectively. Impact from specific filters are included below.

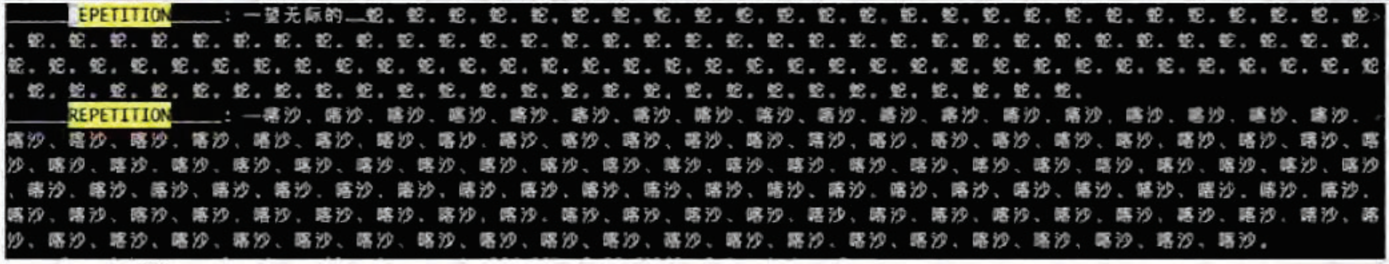
	Fiction	Scitech
_____REPETITION_____	520	242
_____PII_____	31770	21233

_____Copyright_____	52264	30304
Excessive new line characters removed	1097379580	202740904

- Location (RSC):
  - fiction: /[redacted]/fair_llm/[redacted]libgen/[redacted]
  - scitech[redacted]fair_llm[redacted]ibgen/[redacted]
- Examples of filtered data
  - Repetition

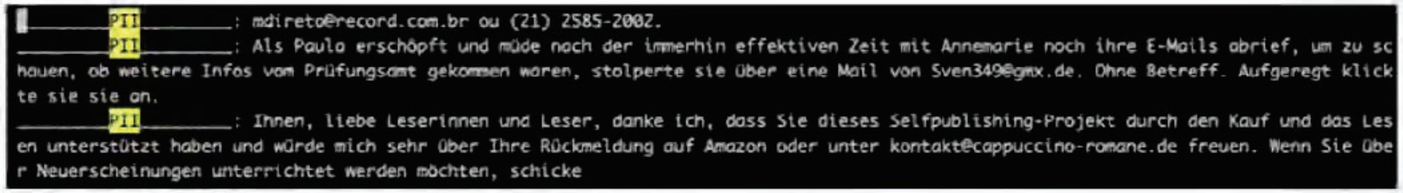


```
______REPETITION______: Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo!
Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo
! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fancu
lo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fan
culo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! F
anculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie!
 Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazi
e! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Gra
zie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! G
razie! Fanculo! Grazie! Fanculo! Grazie! Fanculo! Grazie! No, sul serio, fanculo
```



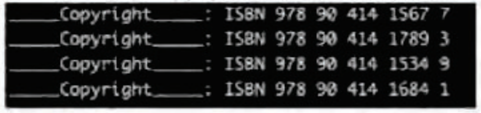
```
______EPETITION______: 一望无际的__蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇>
。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。
蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇
。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。蛇。
______REPETITION______: 一嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙。
嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒
沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙
、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙。
嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒
沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙、嗒沙。
```

  - PII



```
_________PII_________: mdireto@record.com.br ou (21) 2585-2002.
_________PII_________: Als Paula erschöpft und müde noch der immerhin effektiven Zeit mit Annemarie noch ihre E-Mails abrief, um zu sc
hauen, ob weitere Infos vom Prüfungsamt gekommen waren, stolperte sie über eine Mail von Sven349@gmx.de. Ohne Betreff. Aufgeregt klick
te sie sie an.
_________PII_________: Ihnen, liebe Leserinnen und Leser, danke ich, dass Sie dieses Selfpublishing-Projekt durch den Kauf und das Les
en unterstützt haben und würde mich sehr über Ihre Rückmeldung auf Amazon oder unter kontakt@cappuccino-romane.de freuen. Wenn Sie übe
r Neuerscheinungen unterrichtet werden möchten, schicke
```

- Copyright



```
_____Copyright_____: ISBN 978 90 414 1567 7
_____Copyright_____: ISBN 978 90 414 1789 3
_____Copyright_____: ISBN 978 90 414 1534 9
_____Copyright_____: ISBN 978 90 414 1684 1
```

For intermediate output (e.g. what's being filtered), check out the following directories:

- Scitech: [redacted]fair_llm[redacted]ibgen[redacted]
- Fiction: [redacted]fair_llm[redacted]ibgen[redacted]

17.11.2023

LibGen v2

There are a few improvements we can make to LibGen after the manual inspection of the datasets:

- Remove documents, highlighted by the Token Distribution tool: Token Distribution of Training Datasets
- Remove excessive new line character "\n\n\n\n":
  - Limit all the new line characters to 1 "\n"
- Remove repetition:
  - Remove lines that contain <8% unique words, but with at least 100 words
- Remove emails (PII data):

```
- email_regex = re.compile(r'\b[ [illegible] ]{2,}\b')
```

- Remove rows containing copyright in the first and last 25% of the book:
  - Rows containing any of these words: ["ISBN", "Copyright", "©", "All rights reserved", "DOI"]
- [not used] Remove tables of Contents / References / Acknowledgements in the end of the book"

Commented [1]: any rationale of why we're doing this? just better knowledge density? i wonder if it could be useful for long-context?

    - Remove all rows after these words if happen in the last 25% of the document: ["Content", "References", "About the author", "Acknowledgements"]
    - Remove rows with "Content" if happen in the first 10% of the document until the first row that has length more than 30 characters
- [TBD] Split content to Adult/General for LibGen Fiction

Implementation: https://github.com/fairinternal/[redacted]

More details:

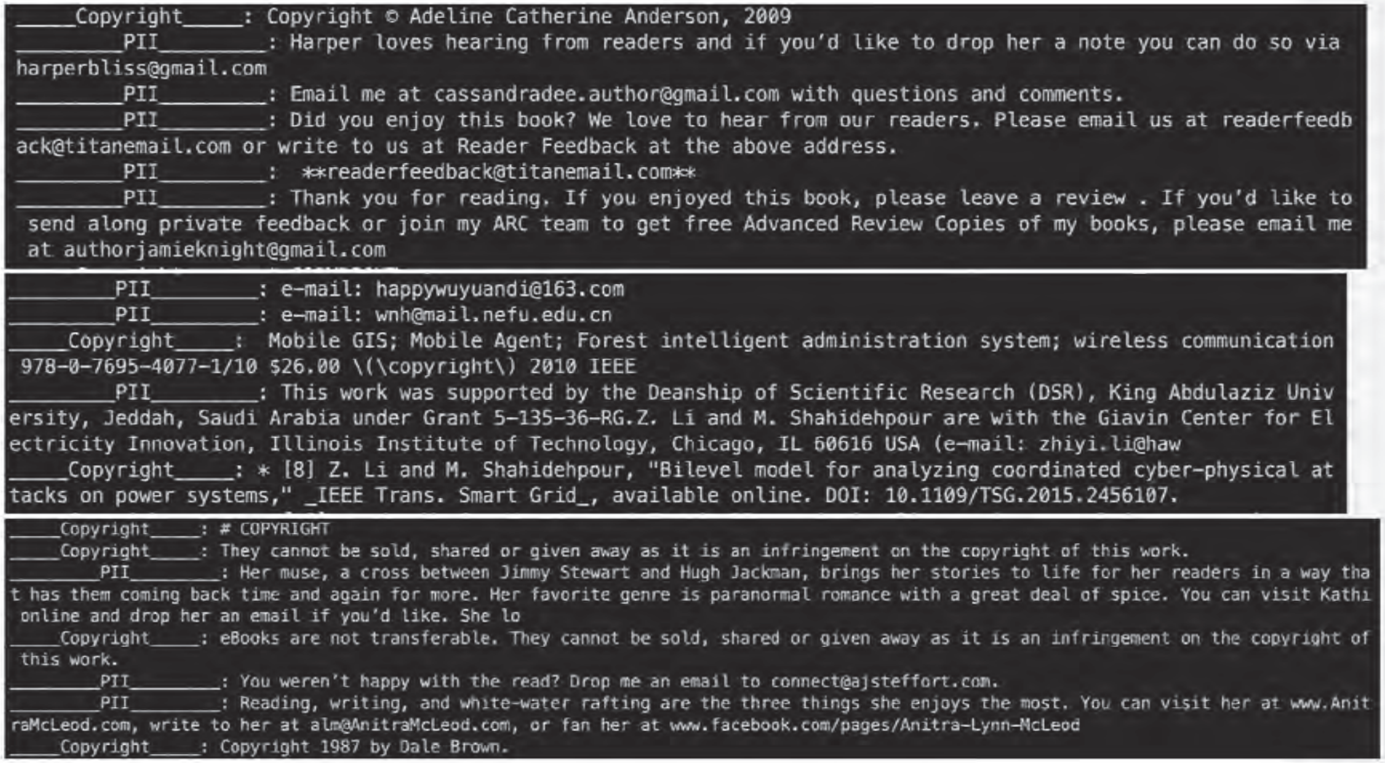
- Observations on LibGen-SciMag
- Data Review: libgen-fiction-books

What was filtered?
We filtered data inside of the documents as well as full documents (based on the Token Distribution outliers):

- Scitech: **0.85%**
- Scimag: **0.28%**
- Fiction: **1.17%**

- scitech: total number of docs: 1255945 | {'lines_copyright_removed': 2334655, 'newlines_removed': 2957148318, 'lines_pii_removed': 1808248, 'lines_repetition_removed': 190613}
- scimag: total number of docs: 41767181 | {'lines_copyright_removed': 16394972, 'newlines_removed': 4191208457, 'lines_pii_removed': 15212651, 'lines_repetition_removed': 410558}
- fiction: total number of docs: 760097 | {'lines_copyright_removed': 125855, 'newlines_removed': 1695675744, 'lines_pii_removed': 101729, 'lines_repetition_removed': 2448}

Copyright&PII (rows removed inside the documents)

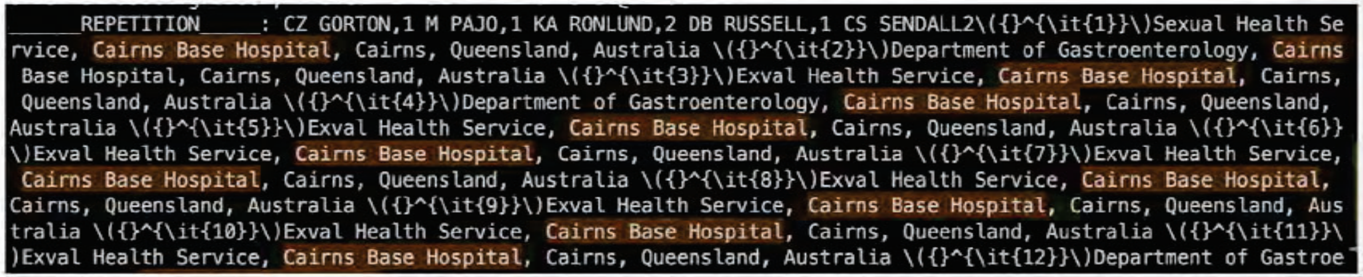


```
_____Copyright_____: Copyright © Adeline Catherine Anderson, 2009
_________PII_________: Harper loves hearing from readers and if you'd like to drop her a note you can do so via
harperbliss@gmail.com
_________PII_________: Email me at cassandradee.author@gmail.com with questions and comments.
_________PII_________: Did you enjoy this book? We love to hear from our readers. Please email us at readerfeedb
ack@titanemail.com or write to us at Reader Feedback at the above address.
_________PII_________:  **readerfeedback@titanemail.com**
_________PII_________: Thank you for reading. If you enjoyed this book, please leave a review . If you'd like to
 send along private feedback or join my ARC team to get free Advanced Review Copies of my books, please email me
 at authorjamieknight@gmail.com
```

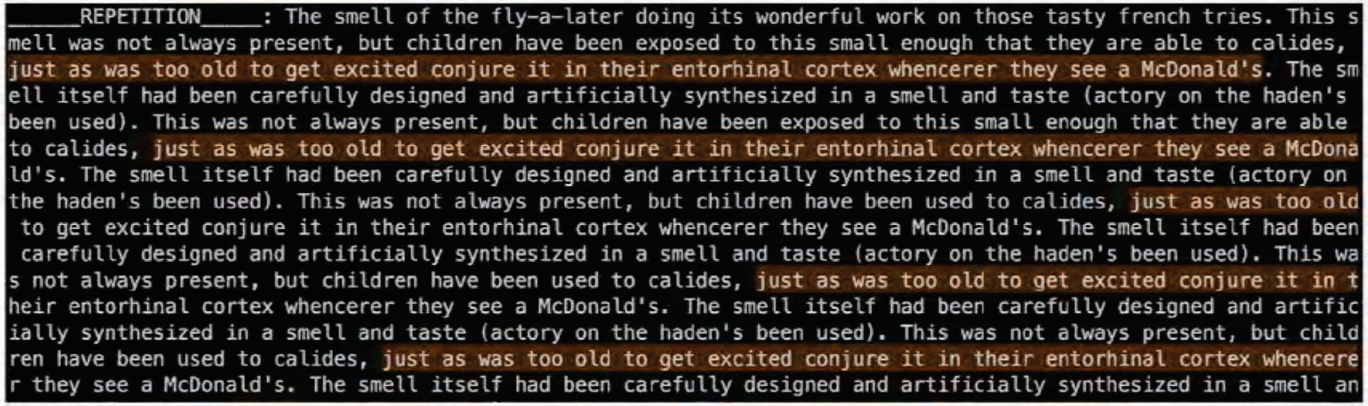
```
_________PII_________: e-mail: happywuyuandi@163.com
_________PII_________: e-mail: wnh@mail.nefu.edu.cn
_____Copyright_____:  Mobile GIS; Mobile Agent; Forest intelligent administration system; wireless communication
 978-0-7695-4077-1/10 $26.00 \(\copyright\) 2010 IEEE
_________PII_________: This work was supported by the Deanship of Scientific Research (DSR), King Abdulaziz Univ
ersity, Jeddah, Saudi Arabia under Grant 5-135-36-RG.Z. Li and M. Shahidehpour are with the Giavin Center for El
ectricity Innovation, Illinois Institute of Technology, Chicago, IL 60616 USA (e-mail: zhiyi.li@haw
_____Copyright_____: * [8] Z. Li and M. Shahidehpour, "Bilevel model for analyzing coordinated cyber-physical at
tacks on power systems," _IEEE Trans. Smart Grid_, available online. DOI: 10.1109/TSG.2015.2456107.
```

```
_____Copyright_____: # COPYRIGHT
_____Copyright_____: They cannot be sold, shared or given away as it is an infringement on the copyright of this work.
_________PII_________: Her muse, a cross between Jimmy Stewart and Hugh Jackman, brings her stories to life for her readers in a way tha
t has them coming back time and again for more. Her favorite genre is paranormal romance with a great deal of spice. You can visit Kathi
 online and drop her an email if you'd like. She lo
_____Copyright_____: eBooks are not transferable. They cannot be sold, shared or given away as it is an infringement on the copyright of
 this work.
_________PII_________: You weren't happy with the read? Drop me an email to connect@ajsteffort.com.
_________PII_________: Reading, writing, and white-water rafting are the three things she enjoys the most. You can visit her at www.Anit
raMcLeod.com, write to her at alm@AnitraMcLeod.com, or fan her at www.facebook.com/pages/Anitra-Lynn-McLeod
_____Copyright_____: Copyright 1987 by Dale Brown.
```

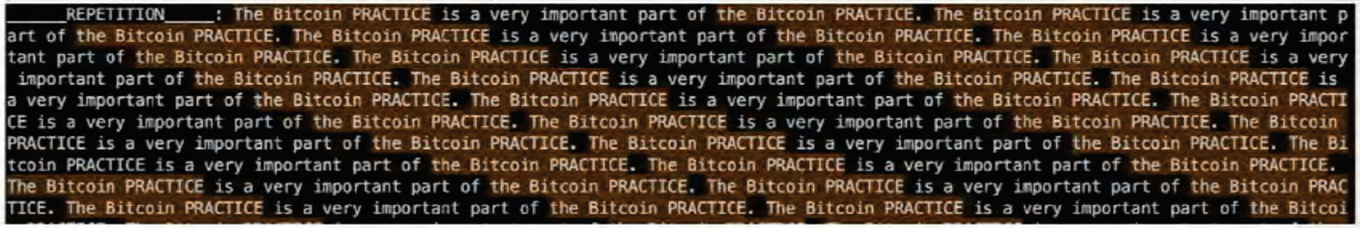
Repetition (Caused by PDF parsing OCR model hallucination. Also removed inside the documents)



```
______REPETITION_____: CZ GORTON,1 M PAJO,1 KA RONLUND,2 DB RUSSELL,1 CS SENDALL2\({}^{\it{1}}\)Sexual Health Se
rvice, Cairns Base Hospital, Cairns, Queensland, Australia \({}^{\it{2}}\)Department of Gastroenterology, Cairns
 Base Hospital, Cairns, Queensland, Australia \({}^{\it{3}}\)Exval Health Service, Cairns Base Hospital, Cairns,
 Queensland, Australia \({}^{\it{4}}\)Department of Gastroenterology, Cairns Base Hospital, Cairns, Queensland,
Australia \({}^{\it{5}}\)Exval Health Service, Cairns Base Hospital, Cairns, Queensland, Australia \({}^{\it{6}}
\)Exval Health Service, Cairns Base Hospital, Cairns, Queensland, Australia \({}^{\it{7}}\)Exval Health Service,
 Cairns Base Hospital, Cairns, Queensland, Australia \({}^{\it{8}}\)Exval Health Service, Cairns Base Hospital,
Cairns, Queensland, Australia \({}^{\it{9}}\)Exval Health Service, Cairns Base Hospital, Cairns, Queensland, Aus
tralia \({}^{\it{10}}\)Exval Health Service, Cairns Base Hospital, Cairns, Queensland, Australia \({}^{\it{11}}\
)Exval Health Service, Cairns Base Hospital, Cairns, Queensland, Australia \({}^{\it{12}}\)Department of Gastroe
```

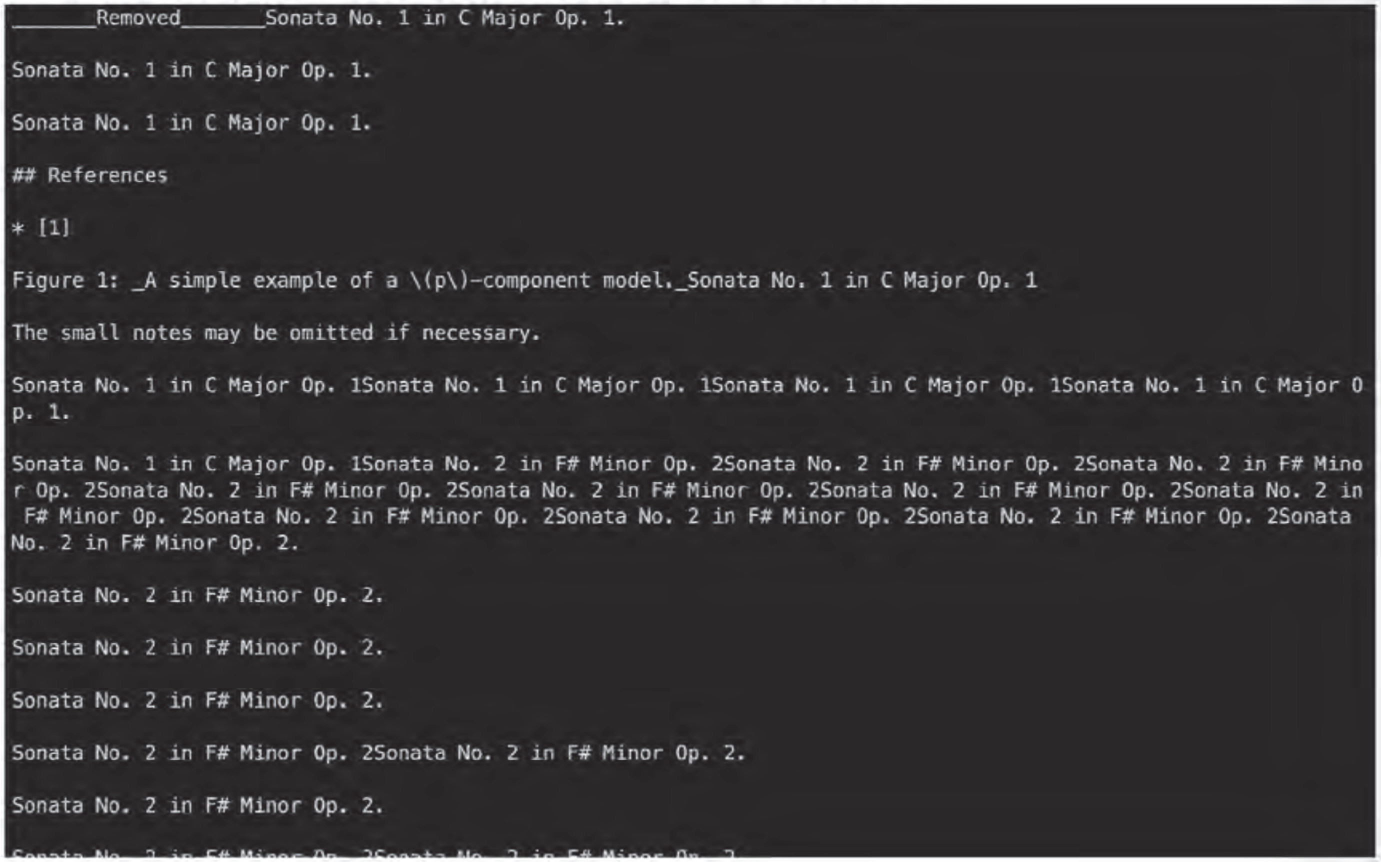


```
______REPETITION_____: The smell of the fly-a-later doing its wonderful work on those tasty french tries. This s
mell was not always present, but children have been exposed to this small enough that they are able to calides,
just as was too old to get excited conjure it in their entorhinal cortex whencerer they see a McDonald's. The sm
ell itself had been carefully designed and artificially synthesized in a smell and taste (actory on the haden's
been used). This was not always present, but children have been exposed to this small enough that they are able
to calides, just as was too old to get excited conjure it in their entorhinal cortex whencerer they see a McDona
ld's. The smell itself had been carefully designed and artificially synthesized in a smell and taste (actory on
the haden's been used). This was not always present, but children have been used to calides, just as was too old
 to get excited conjure it in their entorhinal cortex whencerer they see a McDonald's. The smell itself had been
 carefully designed and artificially synthesized in a smell and taste (actory on the haden's been used). This wa
s not always present, but children have been used to calides, just as was too old to get excited conjure it in t
heir entorhinal cortex whencerer they see a McDonald's. The smell itself had been carefully designed and artific
ially synthesized in a smell and taste (actory on the haden's been used). This was not always present, but child
ren have been used to calides, just as was too old to get excited conjure it in their entorhinal cortex whencere
r they see a McDonald's. The smell itself had been carefully designed and artificially synthesized in a smell an
```

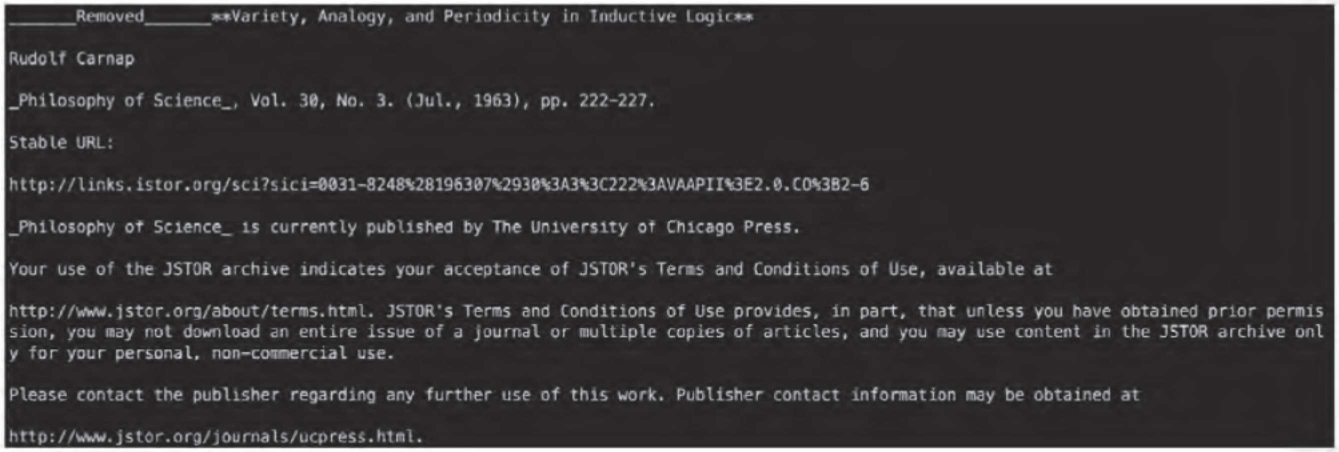


```
______REPETITION_____: The Bitcoin PRACTICE is a very important part of the Bitcoin PRACTICE. The Bitcoin PRACTICE is a very important p
art of the Bitcoin PRACTICE. The Bitcoin PRACTICE is a very important part of the Bitcoin PRACTICE. The Bitcoin PRACTICE is a very impor
tant part of the Bitcoin PRACTICE. The Bitcoin PRACTICE is a very important part of the Bitcoin PRACTICE. The Bitcoin PRACTICE is a very
 important part of the Bitcoin PRACTICE. The Bitcoin PRACTICE is a very important part of the Bitcoin PRACTICE. The Bitcoin PRACTICE is
a very important part of the Bitcoin PRACTICE. The Bitcoin PRACTICE is a very important part of the Bitcoin PRACTICE. The Bitcoin PRACTI
CE is a very important part of the Bitcoin PRACTICE. The Bitcoin PRACTICE is a very important part of the Bitcoin PRACTICE. The Bitcoin
PRACTICE is a very important part of the Bitcoin PRACTICE. The Bitcoin PRACTICE is a very important part of the Bitcoin PRACTICE. The Bi
tcoin PRACTICE is a very important part of the Bitcoin PRACTICE. The Bitcoin PRACTICE is a very important part of the Bitcoin PRACTICE.
The Bitcoin PRACTICE is a very important part of the Bitcoin PRACTICE. The Bitcoin PRACTICE is a very important part of the Bitcoin PRAC
TICE. The Bitcoin PRACTICE is a very important part of the Bitcoin PRACTICE. The Bitcoin PRACTICE is a very important part of the Bitcoi
```

Removed documents (0.25% outliers based on Token Distribution. Removed full documents):



```
_______Removed_______Sonata No. 1 in C Major Op. 1.

Sonata No. 1 in C Major Op. 1.

Sonata No. 1 in C Major Op. 1.

## References

* [1]

Figure 1: _A simple example of a \(p\)-component model._Sonata No. 1 in C Major Op. 1

The small notes may be omitted if necessary.

Sonata No. 1 in C Major Op. 1Sonata No. 1 in C Major Op. 1Sonata No. 1 in C Major Op. 1Sonata No. 1 in C Major O
p. 1.

Sonata No. 1 in C Major Op. 1Sonata No. 2 in F# Minor Op. 2Sonata No. 2 in F# Minor Op. 2Sonata No. 2 in F# Mino
r Op. 2Sonata No. 2 in F# Minor Op. 2Sonata No. 2 in F# Minor Op. 2Sonata No. 2 in F# Minor Op. 2Sonata No. 2 in
 F# Minor Op. 2Sonata No. 2 in F# Minor Op. 2Sonata No. 2 in F# Minor Op. 2Sonata No. 2 in F# Minor Op. 2Sonata
No. 2 in F# Minor Op. 2.

Sonata No. 2 in F# Minor Op. 2.

Sonata No. 2 in F# Minor Op. 2.

Sonata No. 2 in F# Minor Op. 2.

Sonata No. 2 in F# Minor Op. 2Sonata No. 2 in F# Minor Op. 2.

Sonata No. 2 in F# Minor Op. 2.

Sonata No. 2 in F# Minor Op. 2Sonata No. 2 in F# Minor Op. 2
```



```
______Removed______**Variety, Analogy, and Periodicity in Inductive Logic**

Rudolf Carnap

_Philosophy of Science_, Vol. 30, No. 3. (Jul., 1963), pp. 222-227.

Stable URL:

http://links.istor.org/sci?sici=0031-8248%28196307%2930%3A3%3C222%3AVAAPII%3E2.0.CO%3B2-6

_Philosophy of Science_ is currently published by The University of Chicago Press.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at

http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permis
sion, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive onl
y for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at

http://www.jstor.org/journals/ucpress.html.
```

More details [redacted]
Scitech: [redacted] fair_llm [redacted] libger [redacted]
Scimag [redacted] /fair_llm [redacted] /libgen [redacted]
Fiction [redacted] fair_lln [redacted] libger [redacted]

Location [redacted]
fiction [redacted] air_llm [redacted] ibgen [redacted]
scitech [redacted] air_ll [redacted] ibger [redacted]
scimag [redacted] fair_llm [redacted] ibgen/ [redacted]

Ablation experiment:

```
conda activate [redacted] /fair_llm/ [redacted]
```

```
squeue [redacted]o "%.7i %.6P %.30j %.8u %.2t %.6M %.5D %.12q %R"

# monitor job
python -m scripts.monitor [redacted]llama2_libgen_v2_256g_run000 --hanging_timeout_min 15
```

Run command

```
python stool.py run [redacted]lama2_libgen_v2_256g train.py --sweep
sweeps[redacted]fair_use_lib/231121_7B_llama2_libgen_v2_256g.yaml --mem 480 --ncpu 10 --ngpu 8 -ntasks 256 --nodes 32 --
partition learn --anaconda[redacted]fair_llm/[redacted]--qos fair_llm_pretrain --exclude rsclearn[2662] --
launch_restart_dependencies 1
```

Restart command

```
python stool.py relaunch
[redacted]fair_llm/[redacted]lama2_libgen_v2_256g/[redacted]lama2_libgen_v2_256g_run000/2054563 --exclude
rsclearn[2662] --launch_restart_dependencies 4
```

Rerun evals

```
DISABLE_EVALS=False EVALS_PREDICTOR="xlformers" python -m scripts.thib.relaunch_evals --run_dir
"[redacted]fair_llm[redacted]llama2_libgen_v2_256g[redacted]llama2_libgen_v2_256g_run000" --ngpus 8 --batch_size 20
--valid_args default --rebuild
```

Sweeps:

- Run: https://github.com/fairinternal/[redacted]air_use_lib/231121_7B_llama2_libgen_vs_256g[redacted]
- Baseline: https://github.com/fairinternal[redacted]fair_use_lib/231121_7B_llama2_libgen_v1_256g[redacted]

Results:

The new mix shows **improvement on most of the benchmarks**. Low result on **mmlu** could be explained by high volatility of this benchmark (for example, on step **40k the result is 26.62**, which is **1.6 points higher** then on step **50k**)

Caveat: we compare the results for step 42.5k, since at the moment we didn't have more GPUs to complete the training. The "7B Llama2 + LibGen-v1" is the most relevant baseline, as the difference is the version of LibGen + Open Web Math.

Step 42.5k	**7B Llama2 + LibGen v2 + OWM (step 42.5k)**	**7B Llama2 Dill (step 42.5k)**	**7B Llama2 + Libgen-v1 (step 42k)**	**Delta vs Llama2 Dill**	**Delta vs Llama2 Cin + LibGen-v1**
hellaswag.0_shot.acc_char	69.85	68.79	67.65	1.06	2.20
math.4_shot.1_gen.em	1.30	1.68		-0.38	n/a
nq.5_shot.em	17.65	17.40	13.38	0.25	4.27
tqa.5_shot.em	43.78	43.58	40.24	0.20	3.54
piqa.0_shot.acc_char	76.66	76.55	75.41	0.11	1.25
siqa.0_shot.acc_char	47.03	46.21	45.80	0.82	1.23
mmlu.5_shot.macro_avg.acc_char	24.05	24.14	25.96	-0.09	-1.91
human_eval.0_shot.1_gen.em	2.44	1.83	1.83	0.61	0.61
arc_challenge.0_shot.acc_char	40.34	40.26	38.28	0.09	2.06
ppl.code_py	4.06		4.44	n/a	0.39

Step 50k	**7B Llama2 + LibGen v2 + OWM (step 50k)**	**7B Llama2 + Libgen-v1 (step 48k)**	**7B Llama2 + Libgen-v1 (step 51k)**	**Delta vs Llama2 Cin + LibGen-v1 (step 48k)**	**Delta vs Llama2 Cin + LibGen-v1 (step 51k)**
hellaswag.0_shot.acc_char	70.35	67.64	67.95	2.72	2.40

math.4_shot.1_gen.em	1.76				
nq.5_shot.em	18.25	15.32	15.26	2.94	2.99
tqa.5_shot.em	45.50	42.35	40.61	3.15	4.89
piqa.0_shot.acc_char	76.82	74.92	76.50	1.90	0.33
siqa.0_shot.acc_char	47.34	47.19	46.93	0.15	0.41
mmlu.5_shot.macro_avg.acc_char	25.07	25.94	27.36	-0.87	-2.29
human_eval.0_shot.1_gen.em	2.44	2.44	2.44	0.00	0.00
arc_challenge.0_shot.acc_char	40.52	37.68	38.71	2.83	1.80
ppl.code_py	4.02	4.42	4.40	0.40	0.38

Locations:

- 7B Llama2 + LibGen v2 + OWM: [redacted]fair_llm/[redacted]
- 7B Llama2 Dill: [redacted]air_llm/xldumps/az-230913_211008-gpt4tok/az-230913_211008-gpt4tok_run000/eval/0042500
- 7B Llama2 Cin + Libgen-v1:[redacted]ikbash/eval_results/torchx-pci_7b_tok_cl100k_512_4m_with_libgen_v1-[redacted]

14.09.2023

Jacob Xu run minhash deduplication of scitech, fiction and scimag:

LibGen Part	Clean tokens	Minhash deduped	% duplicates removed	Location deduped
Sci-tech EN	220B	190B	15%	[redacted]

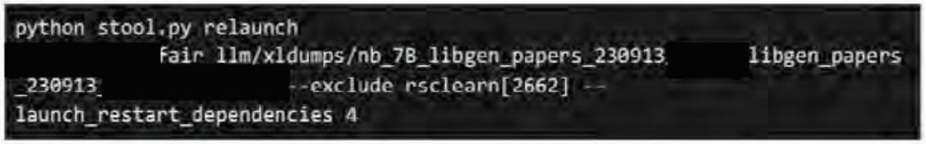
				[redacted]minhashdeduped/lib/sci
Fiction EN	110B	70B	35%	[redacted]nhashdeduped/lib/fic
Sci-mag EN	325B	320B	5%	[redacted]inhashdeduped/lib/sci mag/
Overall	655	560B	**15%**	

13.09.2023

Run ablation experiments for Sci-mag.
Wandb: https://fairwandb.org/fairllm/[redacted]

Targeting 10T datamix with 325B tokens from Sci-mag will make the 1x share of Sci-mag (LibGen papers) to be ~3.3%. To get more signal, we'll assume 2x epochs share in the final datamix, i.e. ~6.5% share.
So the ablation experiment would be to have the Dill datamix + LibGen papers **6.5%** (with reducing proportionally CC share): config.

```
python stool.py run [redacted] libgen_papers_230913 train [redacted] -sweep
[redacted] fair_use_lib/230913_7B_b4M_256gpu_scimag_6pct [redacted] --mem 480 --ncpu 10 --
ngpu 8 --ntasks 256 --nodes 32 --partition learn --anaconda
[redacted] air_llm/envs/[redacted] 0802_pt2 --qos fair_llm --launch_restart_dependencies
4
```



```
python stool.py relaunch
[redacted] fair llm/xldumps/nb_7B_libgen_papers_230913[redacted] libgen_papers
_230913[redacted] --exclude rsclearn[2662] --
launch_restart_dependencies 4
```

26.07.2023

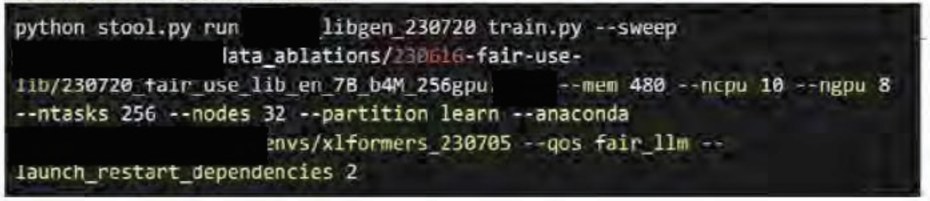
Moved processed sci-mag to S3:

[redacted]/fair-use/scimag/[redacted]

20.07.2023

Re-run LibGen against a new 2k context length Dill baseline datamix:

```
python stool.py run [redacted] libgen_230720 train.py --sweep [redacted] data_ablations/230616-fair-use-
lib/230720_fair_use_lib_en_7B_b4M_256gpu [redacted] --mem 480 --ncpu 10 --ngpu 8 --ntasks 256 --nodes 32 --
partition learn --anaconda [redacted] fair_llm [redacted] --qos fair_llm --
launch_restart_dependencies 2
```



```
python stool.py run [redacted] libgen_230720 train.py --sweep
[redacted] lata_ablations/230616-fair-use-
lib/230720_fair_use_lib_en_7B_b4M_256gpu [redacted] --mem 480 --ncpu 10 --ngpu 8
--ntasks 256 --nodes 32 --partition learn --anaconda
[redacted] envs/xlformers_230705 --qos fair_llm --
launch_restart_dependencies 2
```

Commented [2]: Where are we logging results for this? any more details on the experiment?

Commented [3]: the main results are below (04.07.2023). this was for the new baseline, but we recently changed it to 4k context length, so this run is not relevant (and was stopped).

I will schedule a new run on the new 4k Dill baseline. But we can also use the previous runs (04.07.2023) - they showed positive signals.

Commented [4]:

Commented [5]:

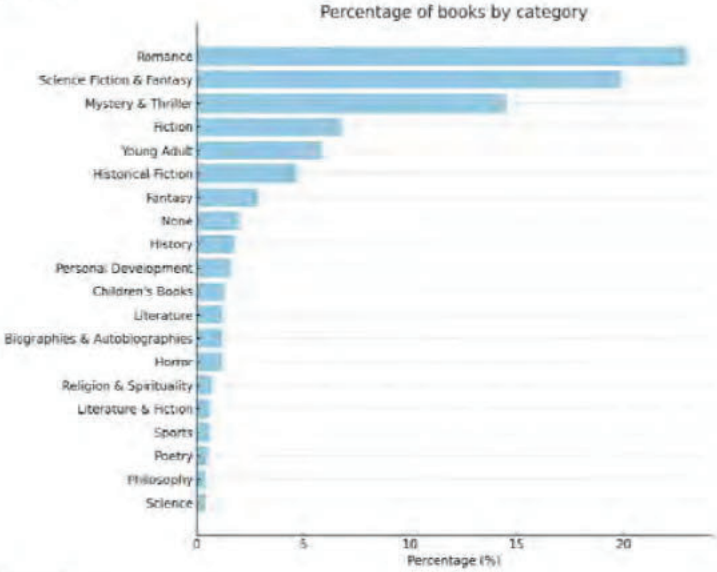
10.07.2023

Re-running evals for mmlu

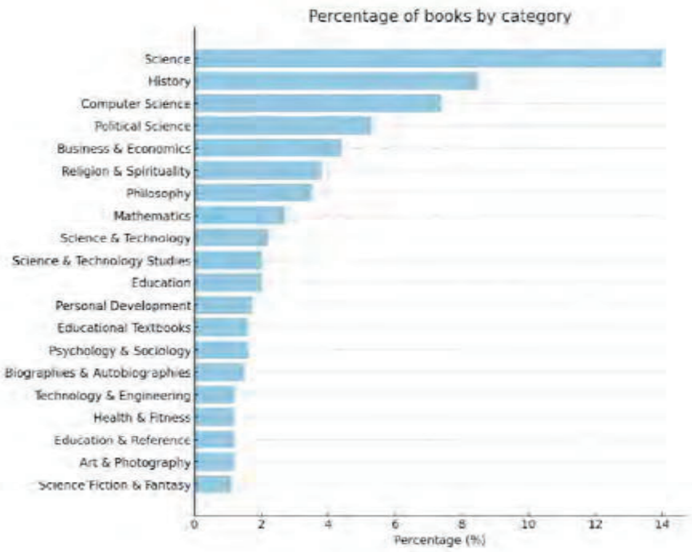
```
python -m scripts.thib.relaunch_evals --run_dir
[redacted]fair_llm/xldumps/[redacted]ibgen_230704/nb_7B[redacted] --ngpus 8 --
batch_size 20 --additional_evals "mmlu"
```

Categorisation of the data (performed by chatLLaMA):

Fiction:



Scitech:



04.07.2023

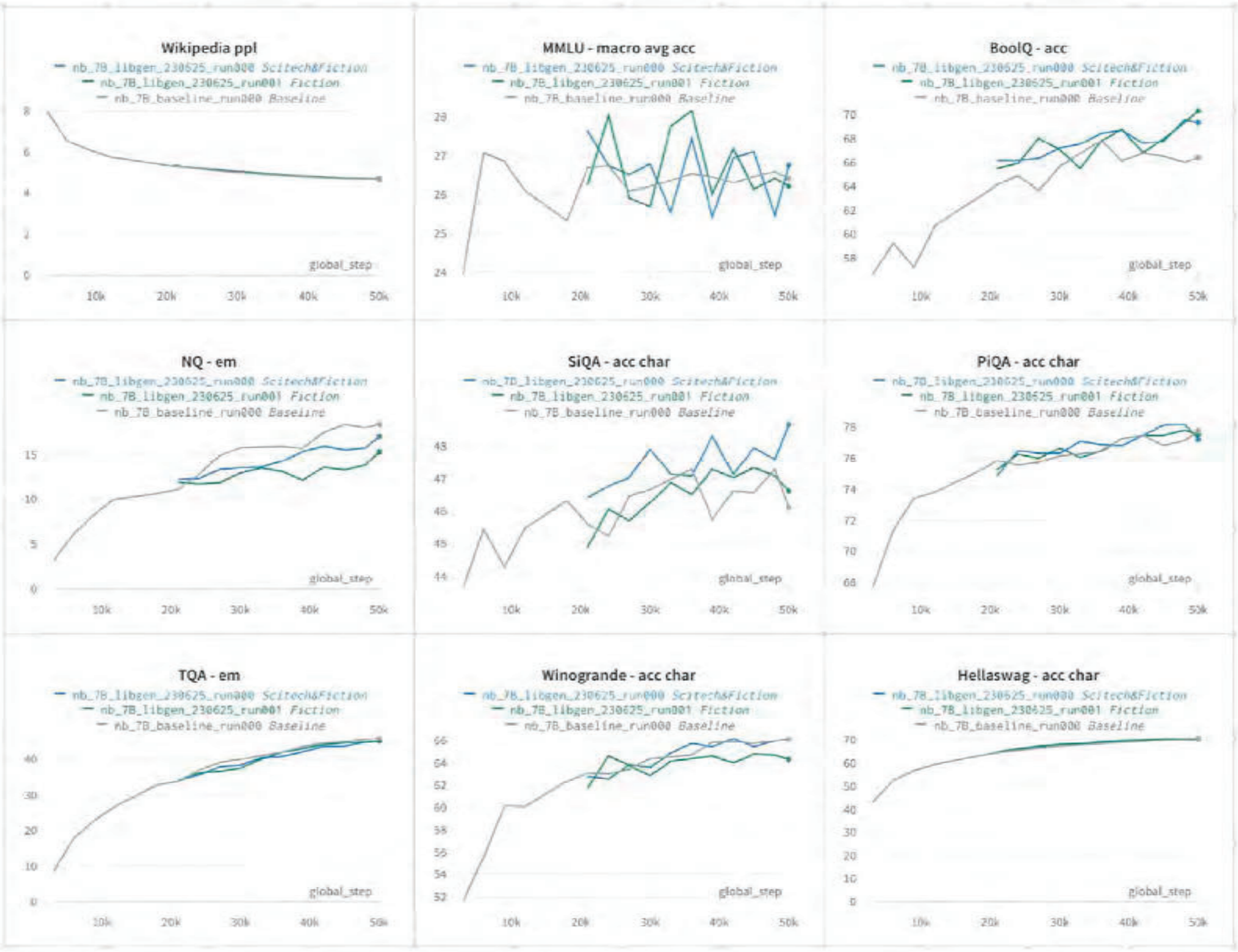
Ablation experiments results:

For experiments **Exp 1** (Scitech+Fiction) and **Exp 2** (Fiction only) we've substituted part of CCNET with LibGen to see the relative impact of the library to the baseline datamix. We observe improvements in the number of metrics:

- **+4.5% BoolQ** (+6% for Exp Fiction only)
- **+5.5% SiQA** (+1.1% for Exp Fiction only)
- **+1.2% MMLU**

Next steps:

- Running **Exp4:** substituting both C4&CCNET with 2 epochs of LibGen. Hypothesis is that we can increase the number of epochs for LibGen
- Running **Exp5**: substituting both C4&CCNET with LibGen in similar proportions. This would be a baseline for **Exp4**


Wikipedia ppl
nb_7B_libgen_230625_run000 Scitech&Fiction
nb_7B_libgen_230625_run001 Fiction
nb_7B_baseline_run000 Baseline
8
6
4
2
0
global_step
10k
20k
30k
40k
50k
MMLU - macro avg acc
nb_7B_libgen_230625_run000 Scitech&Fiction
nb_7B_libgen_230625_run001 Fiction
nb_7B_baseline_run000 Baseline
28
27
26
25
24
global_step
10k
20k
30k
40k
50k
BoolQ - acc
nb_7B_libgen_230625_run000 Scitech&Fiction
nb_7B_libgen_230625_run001 Fiction
nb_7B_baseline_run000 Baseline
70
68
66
64
62
60
58
global_step
10k
20k
30k
40k
50k
NQ - em
nb_7B_libgen_230625_run000 Scitech&Fiction
nb_7B_libgen_230625_run001 Fiction
nb_7B_baseline_run000 Baseline
15
10
5
0
global_step
10k
20k
30k
40k
50k
SiQA - acc char
nb_7B_libgen_230625_run000 Scitech&Fiction
nb_7B_libgen_230625_run001 Fiction
nb_7B_baseline_run000 Baseline
48
47
46
45
44
global_step
10k
20k
30k
40k
50k
PiQA - acc char
nb_7B_libgen_230625_run000 Scitech&Fiction
nb_7B_libgen_230625_run001 Fiction
nb_7B_baseline_run000 Baseline
78
76
74
72
70
68
global_step
10k
20k
30k
40k
50k
TQA - em
nb_7B_libgen_230625_run000 Scitech&Fiction
nb_7B_libgen_230625_run001 Fiction
nb_7B_baseline_run000 Baseline
40
30
20
10
0
global_step
10k
20k
30k
40k
50k
Winogrande - acc char
nb_7B_libgen_230625_run000 Scitech&Fiction
nb_7B_libgen_230625_run001 Fiction
nb_7B_baseline_run000 Baseline
66
64
62
60
58
56
54
52
global_step
10k
20k
30k
40k
50k
Hellaswag - acc char
nb_7B_libgen_230625_run000 Scitech&Fiction
nb_7B_libgen_230625_run001 Fiction
nb_7B_baseline_run000 Baseline
70
60
50
40
30
20
10
0
global_step
10k
20k
30k
40k
50k

- Restarting **Exp 4&5**

nb_7B_libgen_230704_run000

```
python stool.py relaunch
[redacted]'air_llm/xldumps/[redacted]libgen_230704/nb_7B_libgen_230704_[redacted]
[redacted]
rsclearn[1174,1376,1660,2285,2373,2662,2676,2805]
```

nb_7B_libgen_230704_run001

```
python stool.py relaunch
[redacted]air_llm/xldumps[redacted]gen_230704/nb_7B_libgen_230704[redacted]
[redacted]
rsclearn[1174,1376,1660,2285,2373,2662,2676,2805]
```

01.07.2023

Ablation Experiments for LibGen:

- **Exp 1**: Libgen Scitech + Fiction
    - Sweep: https://github.com/fairinternal[redacted]30616-fair-use-lib/230616_fair_use_lib_en_7B_b4M_256[redacted]
    - **Dir exp:**[redacted]**fair_llm/xldumps/**[redacted]
- **Exp 2**: Libgen Fiction only
    - Sweep: https://github.com/fairinternal[redacted]lib/230616_fair_use_lib_en_7B_b4M_256gp[redacted]
    - **Dir exp:**[redacted]**air_llm/xldumps**[redacted]
- **Exp 3**: Libgen vs B3G&Arxiv

  - Sweep: https://github.com/fairinternal[redacted]
  - **Dir exp**[redacted]**ir_llm/xldumps/nb_7B_libgen_230701/**[redacted]
- **Exp 4:** Libgen x2 epochs Scitech+Fiction
  - Sweep: 230704_fair_use_lib_en_7B_b4M_256gp[redacted]
  - **Dir exp:**[redacted]**fair_llm/xldumps/nb_7B_libgen_23070**[redacted]
- **Exp 5:** Libgen vs C4&CCNET
  - Sweep: 230704_fair_use_lib_en_7B_b4M_256gpu.yaml
  - **Dir exp:**[redacted]**fair_llm/xldumps/nb_7B_libgen_230704**[redacted]

Exp 3&5: run command

```
python stool.py run nb_7B_libgen_230704 train.py --sweep
[redacted]230616-fair-use-
lib/230704_fair_use_lib_en_7B_b4M_256gpu[redacted]--mem 480 --ncpu 10 --ngpu 8 --ntasks 256 --
nodes 32 --partition learn --anaconda[redacted]
fair_llm_pretrain --launch_restart_dependencies 2
```

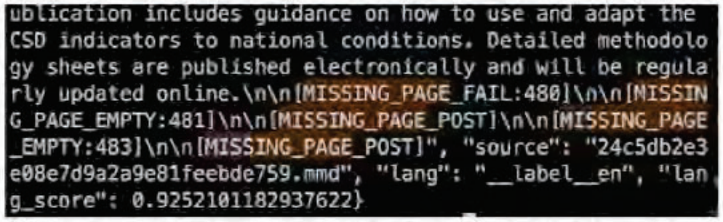
30.06.2023

Statistics on OCR parsing failures:

AVG/doc	fiction_pdf	scitech_pdf	scimag_pdf
num_pages_per_book	170	258	6
num_chars_per_book	344,488	697,960	27,793
num_missing_page_fail_per_book	1.67 page / doc	11.2 page / doc	0.68 page / doc

num_missing_page_post_per_book	0.42 page / doc	14 page / doc	0.05 page / doc
errors_per_char	1.63E-05	7.23E-05	4.21E-05

- Added parsed scitech_pdf and fiction_pdf with markers to determine the page break:
  - Fiction [redacted] ir_llm/data/shuffled/libgen/fiction [redacted]
  - Scitech [redacted] ir_llm/data/shuffled/libgen/scitech [redacted]
  - Marker: "[MISSING_PAGE_*]":
    - MISSING_PAGE_EMPTY
    - MISSING_PAGE_FAIL
    - MISSING_PAGE_POST



```
ublication includes guidance on how to use and adapt the
CSD indicators to national conditions. Detailed methodolo
gy sheets are published electronically and will be regula
rly updated online.\n\n[MISSING_PAGE_FAIL:480]\n\n[MISSIN
G_PAGE_EMPTY:481]\n\n[MISSING_PAGE_POST]\n\n[MISSING_PAGE
_EMPTY:483]\n\n[MISSING_PAGE_POST]", "source": "24c5db2e3
e08e7d9a2a9e81feebde759.mmd", "lang": "__label__en", "lan
g_score": 0.9252101182937622}
```

MISSING_PAGE_EMPTY: (or almost empty) pages. In that case the model tends to collapse into a repetition very quickly. We are catching them at runtime but not always because communication is difficult there. The ones that get through will be caught by the **POST** processing in the very most cases

MISSING_PAGE_FAIL: the model will fail unexplainably somewhere in the page and diverge into a loop. It's determined by a heuristic with a constant threshold so there will be some that will be missed by that. These ones are then caught in the **POST** processing again.

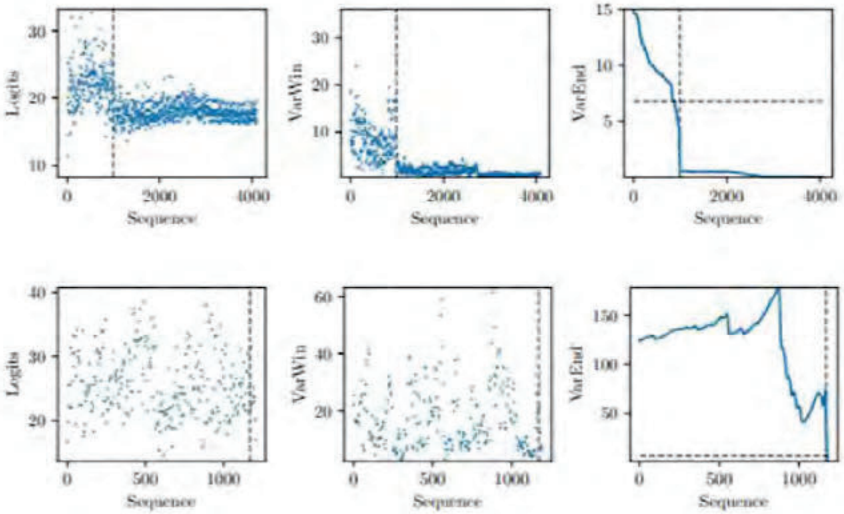


Figure 6: Examples for repetition detection on logits. Top: Sample with repetition, Bottom: Sample without repetition. Left: Highest logit score for each token in the sequence $\ell(x)$, Center: Sliding window variance of the logits $\mathrm{VarWin}_B[\ell](x)$, Right: Variance of variance from the position to the end $\mathrm{VarEnd}_B[\ell](x)$

28.06.2023

[Nikolay]

- Relaunching failed ablation jobs (failed b/c of a bug in the xlformers):
  - ○ [redacted]air_llm/xldumps/nb_7B_libgen_230625/nb_7B_[redacted]
  - ○ [redacted]air_llm/xldumps/nb_7B_libgen_230625/nb_7B[redacted]
- W&B dashboard: https://fairwandb.org/fairllm/[redacted]eports/LibGe[redacted]

```
python stool.py relaunch
[redacted]b_7B_libgen_230701/nb_7B_libgen_230701 [redacted]
```

19.06.2023

Starting an ablation experiment for 100% of EN scitech/fiction (330B tokens). We substitute 10% from CCNet with Libgen scitech dataset (matching it to the target datasets proportion: 2.3T Total vs 330B fiction/scitech -> scitech/fiction is 15%).

Experiments:

- Exp 1: EN: scitech+fiction
  - total dataset: 2.3T tokens
  - scitech&fiction is 330B tokens -> 15%
- Exp 2: EN: fiction
  - total dataset: 2.1T tokens
  - fiction is 110B tokens -> 5%

Dir[redacted]fair_llm/[redacted]

Data:

Data	Total dataset size (billion tokens)	Baseline (weights/ %)	Exp 1 (weights /%)	Exp 2 (weights /%)	Exp 3 (weights /%)	Exp 4 (weights /%)	Exp 5 (weights / %)	Exp 6 (weights / %)	Epochs (# / 200B)

Stack Exchange	25	1.2 (1.8%)						2.2	0.14
B3G (books3 + gutenberg)	28	3 (4.5%)			0			3.6	0.3
Arxiv	33	1.6 (2.4%)			0			2.8	0.15
Github OSS	271	3 (4.5%)						11.6	0.03
C4 en	198	10 (15%)				6 (9%)	7 (10%)	7.7	0.15
CCNet	1,416	45 (67%)	35 (52%)	41.6 (62%)	39.6	29 (43%)	38 (57%)	27.4 + 32.8	E1: 0.07 E2: 0.09
Wikipedia	33	3 (4.5%)						4.3	0.27
Exp 1: Libgen Scitech + Fiction **(nb_7B_libgen_230625_run000)**	330B	-	10 (15%) sci: 6.6 fic: 3.4	-					0.09
Exp 2: Libgen Fiction only **(nb_7B_libgen_2**	110B	-	-	3.4 (5%)					0.09

30625_run001)									
Exp 3: Libgen vs B3G&Arxiv	330B				10 (15%) sci: 6.6 fic: 3.4				0.09
Exp 4: Libgen x2 Scitech+Fiction **(nb_7B_libgen_230704_run000)**	30B					20 (30%) sci: 13.2 fic:6.8			2
Exp 5: Libgen vs C4&CCNET **(nb_7B_libgen_230704_run001)**	30B						10(15%) sci: 6.6 fic: 3.4		1
Exp 6: Libgen - scimag **(nb_7B_libgen_papers_230913_run000)**								scimag: 6.5	
Total	Exp 1,3,5: 2.3T Exp 2: 2.1T Exp 4: 2.7T	67							

Run command (Exp1&2):

```
python stool.py run nb_7B_libgen_230625 train.py --sweep
[redacted]
lib/230616_fair_use_lib_en_7B_b4M_256gpu.yaml --mem 480 --ncpu 10 --ngpu 8 --ntasks 256 --
nodes 32 --partition learn --anaconda [redacted]/fair_llm/[redacted]
fair_llm_pretrain --launch_restart_dependencies 2
```

12-16.06.2023

[Nikolay]

- Total conversion (download -> cleaned):
  - Scitech: 82% (b/c most of scitech are PDFs)
  - Fiction: 86%
  - Scimag: TBD
- non-EN languages:

Language (Sci-tech)	Share, % (Sci-tech)		Language (Fiction)	Share, % (Fiction)
Spanish	23.4%		French	23.1%
Italian	16.0%		German	22.7%
Chinese	13.3%		Spanish	15.0%
Portuguese	11.9%		Dutch	10.5%

German	10.5%		Italian	8.2%
French	7.4%		Hungarian	5.0%
Russian	3.7%		Portuguese	3.5%
Hungarian	2.2%		Chinese	2.8%
Dutch	1.7%		Japanese	2.2%
Turkish	1.1%		Czech	1.5%
Other	8.8%		Other	5.4%

07.06.2023

[Nikolay]

- Added script to convert .mobi to .epub to further parse with epub2markdown script (~60k additional documents, ~10B tokens).
- Converted 7k scitech .mobi to .epub (5% of scitech non-en)

06.06.2023

[Nikolay]

Done with the EN Scitech/Fiction part. Now finishing the non-EN Scitech/Fiction and ALL Scimag.

- Sci-tech (non-en):
    - Downloaded 130k (99%) of non-English epub/mobi Sci-tech books and 586k non-English epub/mobi Fiction books

  - We decided to skip the PDFs for now (since it'll be a hard lift to parse them with our current OCR). There are ~1M non-En PDFs, 60% of which are in Russian (which is not our target language), so the remaining is 425k PDF books (~65B additional multi-lang tokens) which we skip.
- Fiction (non-en):

Libgen Part (non-EN)	Total non-EN (num)	Downloaded (num / %)	Parsed (num / %)	Location Raw	Location Parsed
Sci-tech EPUBs	130,593 (123,281 epub + 7,312 mobi)	128,722 / 99%	0 / 0%	[redacted] air-use/scitech/epub_non_en/	[redacted] fair_llm/data_v2/datasets/books/
Fiction EPUBs	594,348 (545,578 epub + 48,770 mobi)	586,240 / 99%	0 / 0%	[redacted] air-use/fiction/epub_non_en	[redacted] air_llm/data_v2/datasets/books/
Sci-mag All (incl EN)	81,903,411 (876 chunks)	690 / 79%	100 / 11%	[redacted] air-use/scimag/	

05.06.2023

[Nikolay]:

- Sci-mag is 70% downloaded
- Downloaded the remaining 5% of Sci-tech, but all corrupted (unable to parse)
- Parsing multi-lang scitech/fiction PDFs seems to be quite time-consuming - we need to re-train OCR parsing script (no-immediate training data for that), so we'll start with EPUB/MOBI formats for non-English books
- Started loading multi-lang Scitech & Fiction:
    - Fiction (non-en, non-pdf): epub=545,578, mobi=48,770
    - Scitech (non-en, non-pdf): epub=123,281, mobi=7,312
- Convert Scitech multi-lang EPUBs to markdown to check the quality of conversion (could be used for training the OCR for multi-eng)

[Lukas]:

- We can get additional 8-9% of non-English Sci-tech PDFs (~400k books). But for that we need training data for Spanish, German, Italian, French (optional: Chinese, Portuguese):
    - Check if we have training data on Arxiv

Libgen Part (EN)	Total EN (num)	Downloaded (num / %)	Parsed (num / %)	Location Raw	Location OCR Parsed	Location Cleaned	Cleaned tokens (#)

Sci-tech PDFs	1,272,655	1,241,150 / 98%	1,104,047 / 89%	[redacted] air-use/scitech/pdf_en/	/large_experiments/fair_llm/data_v2/ [redacted] fair-use/scitech/pdf_ocr/	[redacted] air-use/scitech/processed	170B
Sci-tech EPUBs	454,064	454,534 / 100%	392,426 / 86%	[redacted] fair-use/scitech/epub_en/	[redacted] air_llm/data_v2/	[redacted] air-use/scitech/processed/202	50B
Fiction PDFs	117,980	106,362/ 90%	96,981 / 82%	[redacted] ir-use/fiction/pdf_en/	[redacted] air_llm/data_v2/datasets/books fair-use/fiction/pdf_ocr/	[redacted] fair-use/fiction/processed/	10B
Fiction EPUBs	1,041,740	1,031,934 / 99%	946,144 / 91%	[redacted] ir-use/fiction/epub_en/	[redacted] air_llm/data_v2/	[redacted] /fair-use/fiction/processed/	100B
Sci-mag	81,903,411 (876 chunk	619 / 70%	100 / 11%	[redacted] fair-	[redacted] fair-use/scimag/pdf_ocr/		

All	s)			use/scimag /			

01.06.2023

Libgen Part (EN)	Total EN (num)	Downloaded (num / %)	Parsed (num / %)	Location Raw	Location Parsed	Location Cleaned	Cleaned tokens (#)
Sci-tech PDFs	1,272,655	1,165,867 / 92%	1,103,695 / 95%	[redacted] fair-use/scitech /pdf_en/	[redacted] ir_llm/data_v2/datasets/books/dat	[redacted] fair- [redacted] scitech/processed /20230526_pdf_en/	170 B
Sci-tech EPUBs	454,064	454,534 / 100%	392,426 / 86%	[redacted] fair- [redacted] citech /epub_en/	[redacted] ir_llm/data_v2/datasets/books/dat	[redacted] fair-use/ scitech/processed/20230526_epub_en/	50B
Fictio	117,980	105,077/	96,981 /	[redacted]	[redacted] air_llm/data_v2/datasets/books/dat	[redacted] air-	10B

n PDFs		89%	82%	data/fair-use/fiction/pdf_en/		use/fiction/processed/20230526_pdf_en/	
Fiction EPUBs	1,041,740	1,022,914 / 98%	946,144 / 91%	[redacted] air-use/fiction/epub_en/	[redacted] ir_llm/data_v2/datasets/books/dat	[redacted] ir-use/fiction/processed/20230526_epub_en/	100 B
Sci-mag All	81,903,411 (876 chunks)	451 / 51%	24 / 3%	[redacted] /fair-use/scimag			

[Lukas]

- Started with Scimag
- Optimized Nougat OCR inference code for many small documents

30.05.2023

Libgen Part (EN)	Total EN (num)	Downloaded (num / %)	Parsed (num / %)	Location Raw	Location Parsed	Location Cleaned	Cleaned tokens (#)

Sci-tech PDFs	1,272,655	1,165,867 / 92%	1,066,478 / 91%	[redacted] [redacted] [redacted]air-use/scitech/pdf_en/	air_llm/data_v2/datasets/books/dat [redacted]	[redacted] [redacted]r-use/scitech/processed/20230526_pdf_en/	170B
Sci-tech EPUBs	454,064	454,534 / 100%	392,426 / 86%	[redacted] [redacted] [redacted]air-use/scitech/epub_en/	ir_llm/data_v2/datasets/books/dat [redacted]	[redacted] [redacted]ir-use/scitech/processed/20230526_epub_en/	50B
Fiction PDFs	117,980	105,077/ 89%	96,981 / 82%	[redacted] [redacted] [redacted]fair-use/fiction/pdf_en/	air_llm/data_v2/datasets/books/dat [redacted]	[redacted] [redacted]ir-use/fiction/processed/20230526_pdf_en/	10B
Fiction EPUBs	1,041,740	1,022,914 / 98%	946,144 / 91%	[redacted] [redacted] [redacted]fair-use/fiction/epub_en/	fair_llm/data_v2/datasets/books/dat [redacted]	[redacted] [redacted]ir-use/fiction/processed/20230526_epub_en/	100B
Sci-mag	81,903,411 (876 chunk	435 / 50%	0	[redacted]			

All	s)			data/fair-use/scimag			

26.05.2023

As of 5pm

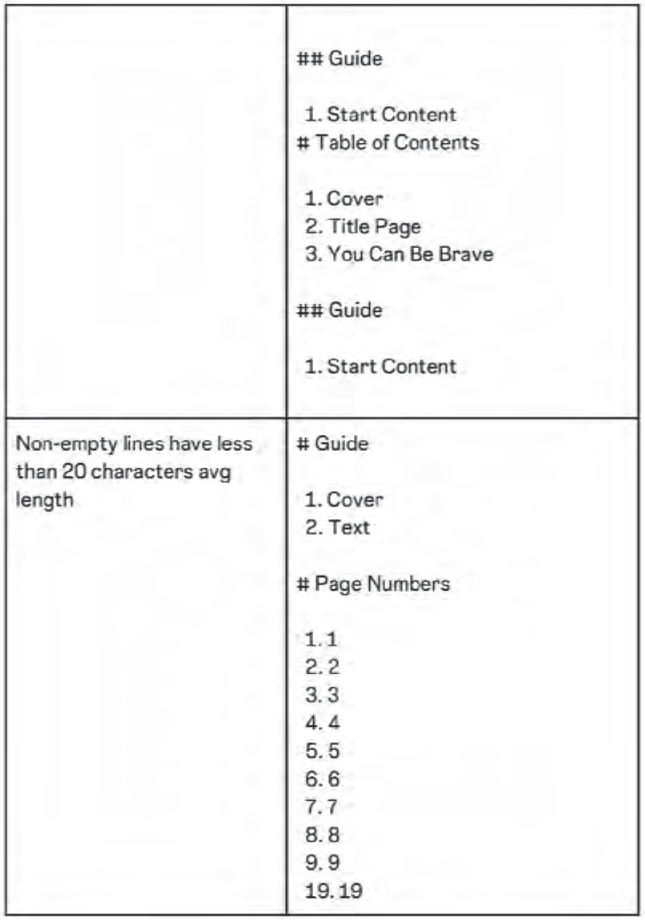
Libgen Part (EN)	Total EN (num)	Downloaded (num / %)	Parsed (num / %)	Location Raw	Location Parsed	Location Cleaned	Cleaned tokens (#)
Sci-tech PDFs	1,272,655	1,165,867 / 92%	1,066,478 / 91%	[redacted] air-use/scitech/pdf_en/	[redacted] fair_llm/data_v2/datasets/books/dat [redacted]	[redacted] fair-use/scitech/processed/20230526_pdf_en/	170B
Sci-tech EPUBs	454,064	454,534 / 100%	392,426 / 86%	[redacted] data/fair-use/scitech/epub_en/	[redacted] fair_llm/data_v2/datasets/books/dat [redacted]	[redacted] air-use/scitech/processed/20230526_epub_en/	50B

Fiction PDFs	117,980	105,077/ 89%	96,981 / 82%	[redacted]air-use/fiction/pdf_en/	[redacted] fair_llm/data_v2/datasets/books/dat [redacted]	[redacted]ir-use/fiction/processed/20230526_pdf_en/	10B
Fiction EPUBs	1,041,740	1,022,914 / 98%	946,144 / 91%	[redacted]ir-use/fiction/epub_en/	[redacted] fair_llm/data_v2/datasets/books/dat [redacted]	[redacted]air-use/fiction/processed/20230526_epub_en/	100B
Sci-mag All	81,903,411 (876 chunks)	361 / 41%	0	[redacted]fair-use/scimag			

[Peter]

- Also had memory limitations
- Finalized book filters:

Condition	Example of an affected book
Book line count less than 50	# Table of Contents 1. Cover 2. Title Page 3. You Can Be Brave



	## Guide 1. Start Content # Table of Contents 1. Cover 2. Title Page 3. You Can Be Brave ## Guide 1. Start Content
Non-empty lines have less than 20 characters avg length	# Guide 1. Cover 2. Text # Page Numbers 1. 1 2. 2 3. 3 4. 4 5. 5 6. 6 7. 7 8. 8 9. 9 19. 19

	20. 20
	21. 21
	22. 22
	23. 23
	24. 24
	25. 25
	26. 26

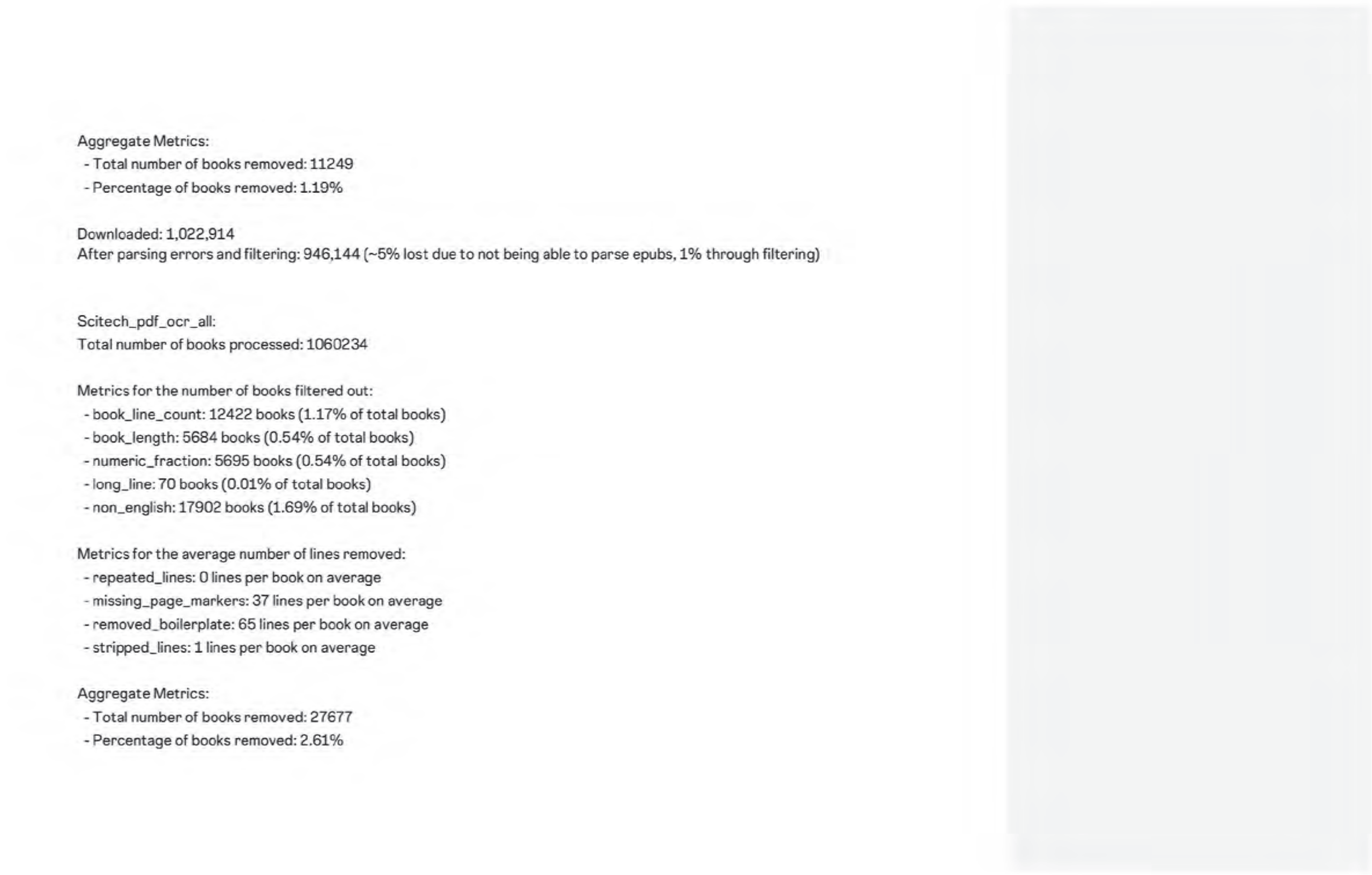
Numeric fraction of characters > 10%	1. 2
	2. 3
	3. 4
	4. 5
	5. 6
	6. 7
	16. 17
	17. 18
	18. 19
	19. 20
	20. 21

Line longer than 50k characters	Book without any new lines or formatting, sometimes a parsing issue
Language id less than 0.5 for english	Our pdf ocr model is trained on english documents, so there are hallucinations when ocring non-english text. Also we only want english book for

	now. P.- J. HÉRAULT CAL DE TER COLLECTION « ANTICIPATION » ÉDITIONS FLEUVE NOIR 6, rue Garantière – PARIS VIe

- Scimag [redacted]air-use/scimag

```
Stat for filtering Fiction_epub:

Total number of books processed: 945531

Metrics for the number of books filtered out:
 - book_line_count: 4951 books (0.52% of total books)
 - book_length: 1928 books (0.20% of total books)
 - numeric_fraction: 261 books (0.03% of total books)
 - long_line: 3362 books (0.36% of total books)
 - non_english: 5602 books (0.59% of total books)

Metrics for the average number of lines removed:
 - repeated_lines: 0 lines per book on average
 - missing_page_markers: 0 lines per book on average
 - removed_boilerplate: 97 lines per book on average
 - stripped_lines: 4 lines per book on average
```



Aggregate Metrics:

- Total number of books removed: 11249
- Percentage of books removed: 1.19%

Downloaded: 1,022,914
After parsing errors and filtering: 946,144 (~5% lost due to not being able to parse epubs, 1% through filtering)

Scitech_pdf_ocr_all:
Total number of books processed: 1060234

Metrics for the number of books filtered out:

- book_line_count: 12422 books (1.17% of total books)
- book_length: 5684 books (0.54% of total books)
- numeric_fraction: 5695 books (0.54% of total books)
- long_line: 70 books (0.01% of total books)
- non_english: 17902 books (1.69% of total books)

Metrics for the average number of lines removed:

- repeated_lines: 0 lines per book on average
- missing_page_markers: 37 lines per book on average
- removed_boilerplate: 65 lines per book on average
- stripped_lines: 1 lines per book on average

Aggregate Metrics:

- Total number of books removed: 27677
- Percentage of books removed: 2.61%

25.05.2023

[Nikolay] Had memory limitation on fair cluster of (20T) so had to back up everything to s3:

- Fiction: [redacted]air-use/fiction
- Scitech [redacted]/fair-use/scitech
- Scimag: [redacted]air-use/scimag

24.05.2023

[Lukas]

- Script to filter SciMag files (script):
  - Checks if file is corrupt
  - Checks if file is PDF
  - Checks if PDF text is english

  → Send to Nougat OCR

As of 5pm

Libgen Part (EN)	Total EN (num)	Downloaded (num / %)	Parsed (num / %)	Location Raw	Location Parsed	Location Cleaned
Sci-tech PDFs	1,272,665	1,072,286 / 84%	1,025,070 / 81%	[redacted] fair-use/scitech/en_pdf	[redacted]air_llm/data_v2/datasets/books/[redacted]	

Sci-tech EPUBs	454,064	454,534 / 100%	392,426 / 86%	[redacted] fair-use/scitech/en_epub		
Fiction PDFs	117,980	102,118/ 87%	96,981 / 82%	[redacted] air-use/fiction/en_pdf	[redacted] fair_llm/data_v2/datasets/books/data/ [redacted]	
Fiction EPUBs	1,041,740	1,001,538 / 96%	642,703 / 64%	[redacted] air-use/fiction/en_epub		
Sci-mag All	81,903,411	37,713 / 0%	0	[redacted] air-use/scimag		

23.05.2023

Notes:

- Trying to load scimag with the same approach (direct download) as before - doesn't seem to be fast (250k docs / 12h -> 160 days to download the library). Exploring other options to load faster.

As of 5pm

Libgen	Total EN (num)	Downloaded (num /	Parsed (num /	Location Raw	Location Parsed

Part (EN)		%)	%)		
Sci-tech PDFs	1,272,655	1,179,045 / 93%	973,340 / 76%	[redacted] ir- [redacted]tech/en_pdf	[redacted]air_llm/data_v2/datasets/books/[redacted] [redacted]
Sci-tech EPUBs	454,064	454,534 / 100%	392,426 / 86%	[redacted] air- use/scitech/en_epub	[redacted]scitech_[redacted]
Fiction PDFs	117,980	102,118/ 87%	69,423 / 59%	[redacted] air- use/fiction/en_pdf	[redacted]air_llm/data_v2/datasets/books/[redacted] [redacted]
Fiction EPUBs	1,041,740	1,001,538 / 96%	642,703 / 64%	[redacted] air- use/fiction/en_epub	[redacted]data/fiction_[redacted]
Sci-mag All	81,903,411	0	0	[redacted] air- use/scimag	

22.05.2023

Notes:

- On the weekend hit the hard limit of disk utilization on fair cluster: ~24T (in my personal folder nikbash)
- Had to clean the disk (what was possible to clean), so now around ~21T
- With these constraints can't easily load scimag (~80T), so
    - EITHER distribute download across team (we have same IP, so would be throttled by libgen)
    - OR transfer raw files to S3, remove them from fair cluster (need to finish processing first) and load scimag in chunks
    - OR increase the disk space
- The problem with scimag loading is that there is no metadata for it, so we can't pre-filter by language and extension first, so we need to load everything at once (in chunks)
- Started backing up raw data to S3 bucket (to further remove raw data from the fair cluster)
    - Fiction:
        - EPUBs: [redacted] /fair-use/fiction/epub_en/
        - PDFs: [redacted] air-use/fiction/pdf_en/
    - Scitech:
        - EPUBs: [redacted] air-use/scitech/epub_en/
        - PDFs: [redacted] air-use/scitech/pdf_en/

As of 5pm

Libgen Part (EN)	Total EN (num)	Downloaded (num / %)	Parsed (num / %)	Location Raw	Location Parsed
Sci-tech PDFs	1,272,655	1,179,045 / 93%	904,149 / 71%	[redacted] [redacted] ir-use/scitech/en_pdf	[redacted] ir_llm/data_v2/datasets/books/ [redacted] [redacted]
Sci-tech EPUBs	454,064	454,534 / 100%	392,426 / 86%	[redacted] [redacted] air-	[redacted] ata/scitech [redacted]

				use/scitech/en_epub	
Fiction PDFs	117,980	102,118/ 87%	67,192 / 57%	[redacted] ir- [redacted] tion/en_pdf	fair_llm/data_v2/datasets/books/[redacted]
Fiction EPUBs	1,041,740	1,001,538 / 96%	642,703 / 64%	[redacted] use/fiction/en_epub	[redacted] data/fiction_[redacted]
Sci-mag All	81,903,411	0	0	[redacted] ir- use/scimag	

19.05.2023

Notes:

- The download speed dropped significantly for the remaining 15% of data (probably the data is on the servers with low throughput)
- Planning to start loading Sci-mag on the weekend
- Discussed with Lukas Blecher that we would need to train the Nougat OCR on other languages to be able to parse the non-EN PDFs somewhere around end of June, 23
- Started parsing Fiction PDFs with Nougat OCR:
    - The quality of other PDF parsers was not satisfactory (see notes from 18.05.2023)
    - The number of EN PDFs in Fiction is relatively small - 117k, so we need just 1-2 days with 500 GPUs

As of 5pm

Libgen Part (EN)	Total EN (num)	Downloaded (num / %)	Parsed (num / %)	Location Raw	Location Parsed
Sci-tech PDFs	1,272,655	1,006,428 / 80%	822,167 / 65%	[redacted] fair-use/scitech/en_pdf	[redacted] air_llm/data_v2/datasets/books [redacted]
Sci-tech EPUBs	454,064	454,356 / 100%	392,426 / 86%	[redacted] air-use/scitech/en_epub	[redacted] citech [redacted]
Fiction PDFs	117,980	69,554 / 59%	0	[redacted] /fair-use/fiction/en_pdf	
Fiction EPUBs	1,041,740	780,513 / 78%	642,703 / 64%	[redacted] air-use/fiction/en_epub	[redacted] iction_ [redacted]
Sci-mag All	81,903,411	0	0	[redacted] air-use/scimag	

18.05.2023

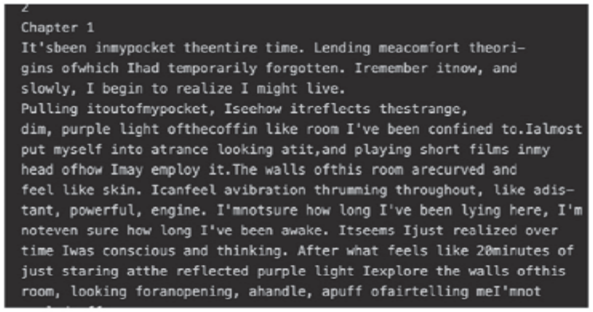
As of 5pm

Libgen Part (EN)	Total EN (num)	Downloaded (num / %)	Parsed (num / %)	Location Raw	Location Parsed
Sci-tech PDFs	1,272,655	1,006,428 / 80%	645,061 / 51%	[redacted] [redacted]air-use/scitech/en_pdf	[redacted]air_llm/data_v2/datasets/books[redacted] [redacted]
Sci-tech EPUBs	454,064	454,292 / 100%	392,426 / 86%	[redacted] [redacted]air-use/scitech/en_epub	[redacted]/libgen_epub[redacted]
Fiction PDFs	117,980	67,274 / 50%	0	[redacted] [redacted]air-use/fiction/en_pdf	
Fiction EPUBs	1,041,740	696,542 / 70%	642,703 / 64%	[redacted] [redacted]air-use/fiction/en_epub	[redacted]data/fiction[redacted]
Sci-mag All	81,903,411	0	0	[redacted] [redacted]air-use/scimag	

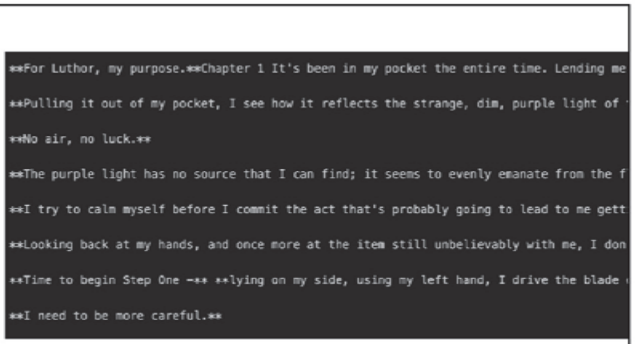
Notes:

- We looked at processing the Fiction PDFs with a non-ocr parser PYPDF2, as it would be much faster. But even with normal novels there are lots of artifact like missing spaces or random spaces within words
- Therefore we decided to also use nougat ocr for all the fiction pdfs

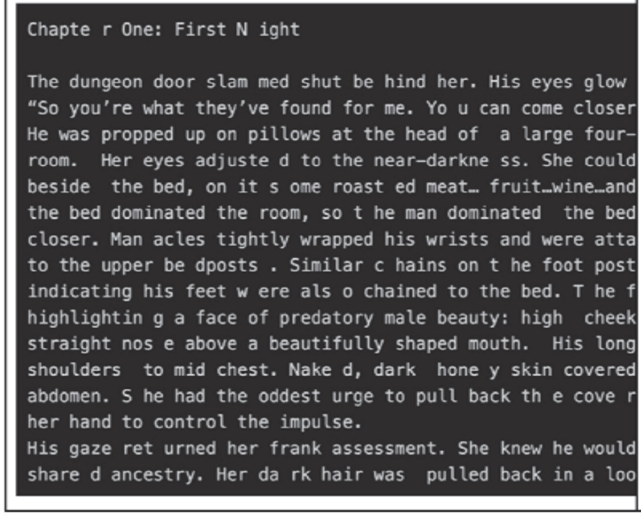
PYDF2 (with spacing issues)	Nougat OCR



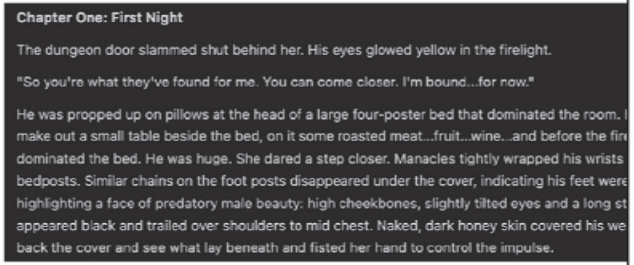
```
2
Chapter 1
It'sbeen inmypocket theentire time. Lending meacomfort theori-
gins ofwhich Ihad temporarily forgotten. Iremember itnow, and
slowly, I begin to realize I might live.
Pulling itoutofmypocket, Iseehow itreflects thestrange,
dim, purple light ofthecoffin like room I've been confined to.Ialmost
put myself into atrance looking atit,and playing short films inmy
head ofhow Imay employ it.The walls ofthis room arecurved and
feel like skin. Icanfeel avibration thrumming throughout, like adis-
tant, powerful, engine. I'mnotsure how long I've been lying here, I'm
noteven sure how long I've been awake. Itseems Ijust realized over
time Iwas conscious and thinking. After what feels like 20minutes of
just staring atthe reflected purple light Iexplore the walls ofthis
room, looking foranopening, ahandle, apuff ofairtelling meI'mnot
```



```
**For Luthor, my purpose.**Chapter 1 It's been in my pocket the entire time. Lending me

**Pulling it out of my pocket, I see how it reflects the strange, dim, purple light of

**No air, no luck.**

**The purple light has no source that I can find; it seems to evenly emanate from the f

**I try to calm myself before I commit the act that's probably going to lead to me gett

**Looking back at my hands, and once more at the item still unbelievably with me, I don

**Time to begin Step One -** **lying on my side, using my left hand, I drive the blade

**I need to be more careful.**
```



```
Chapte r One: First N ight

The dungeon door slam med shut be hind her. His eyes glow
“So you’re what they’ve found for me. Yo u can come closer
He was propped up on pillows at the head of  a large four-
room.  Her eyes adjuste d to the near-darkne ss. She could
beside  the bed, on it s ome roast ed meat… fruit…wine…and
the bed dominated the room, so t he man dominated  the bed
closer. Man acles tightly wrapped his wrists and were atta
to the upper be dposts . Similar c hains on t he foot post
indicating his feet w ere als o chained to the bed. T he f
highlightin g a face of predatory male beauty: high  cheek
straight nos e above a beautifully shaped mouth.  His long
shoulders  to mid chest. Nake d, dark  hone y skin covered
abdomen. S he had the oddest urge to pull back th e cove r
her hand to control the impulse.
His gaze ret urned her frank assessment. She knew he would
share d ancestry. Her da rk hair was  pulled back in a loo
```



Chapter One: First Night

The dungeon door slammed shut behind her. His eyes glowed yellow in the firelight.

"So you're what they've found for me. You can come closer. I'm bound...for now."

He was propped up on pillows at the head of a large four-poster bed that dominated the room. I
make out a small table beside the bed, on it some roasted meat...fruit...wine...and before the fir
dominated the bed. He was huge. She dared a step closer. Manacles tightly wrapped his wrists
bedposts. Similar chains on the foot posts disappeared under the cover, indicating his feet were
highlighting a face of predatory male beauty: high cheekbones, slightly tilted eyes and a long st
appeared black and trailed over shoulders to mid chest. Naked, dark honey skin covered his we
back the cover and see what lay beneath and fisted her hand to control the impulse.

17.05.2023

As of 5pm

Libgen Part	Total EN (num)	Downloaded (num / %)	Parsed (num / %)	Location Raw	Location Parsed
Sci-tech PDFs	1,272,655	835,499 / 65%	645,061 / 51%	[redacted]/libgen_pdf	[redacted]air_llm/data_v2/datasets/books/[redacted]
Sci-tech EPUBs	454,064	454,292 / 100%	0	[redacted]/libgen_epub	[redacted]data/libgen_[redacted]
Fiction PDFs	117,980	58,071 / 49%	0	[redacted]fiction/fiction_pdf	
Fiction EPUBs	1,041,740	627,218 / 60%	0	[redacted]/fiction/fiction_epub	
Sci-mag	81,903,411	0	0		

All					

16.05.2023

[Lukas]

Sci-Tech conversion status (6pm 16.05.2023): (38% done of 1,726,719)

- PDFs (579,620 or 46% of 1,272,655): [redacted] air_llm/data_v2/datasets/books/[redacted]
- EPUBs (82,699 or 18% of 454,064): [redacted] data/libgen_[redacted]

[Nikolay]

Scitech EN download status (6pm 16.05.2023): (95% done of 1,201,994)

Fiction EN download status (6pm 16.05.2023): (55% done of 1,159,720)

- EPUBs (580,899): [redacted] ction/fiction_epub
- PDFs (55,633) : [redacted] iction/fiction_pdf

Robert Stojnic suggested that we could do an experiment with finetuning 70B model on the sci-tech data to check that it would improve the reasoning capabilities (ideally to match the Galactica):

Option	GPU hours	Comment
70B on 512 GPUs	362h (15 days)	sci-tech tokens (1 epoch): 200B wps 70B: 300 GPU*hours = 200B/(num_g*wps*3600s)
70B on 1024 GPUs	181h (7.5 days)	
70B on 2048 GPUs	90h (3.7 days)	

15.05.2023

[Lukas]

Sci-Tech conversion status (5pm 15.05.2023): (34% done of 1,726,719)

- PDFs (499,404 or 39% of 1,272,655): [redacted] air_llm/data_v2/datasets/books/[redacted]
- EPUBs (82,699 or 18% of 454,064): [redacted] ata/libgen_epub_parsed

SciMag calculation:

Processing speed: (12.6±10.5) s/batch @ 4 pages per batch
#pages SciMag: 50%*82M*6=246M pages (assume 50% english)
Estimated GPU hours: (12.6±10.5)*246M/4/3600= (215±180)k GPUh

[Nikolay]

Instructions to download libgen:

- [redacted] fair_data/fair_data/projects/fair_use_lib
- in your fair cluster terminal run "screen -S fiction"
- in a new screen window:
  - source activate [redacted]
  - [redacted] ibgen_direct.py"

Scitech EN download status (12pm 15.05.2023): (95% done of 1,201,994)

- EPUBs (310k):
  - 111,292 on FAIR Cluster: [redacted] bgen_epub
  - 199,145 on [redacted]
- PDFs (847k):
  - 647,932 on Fair Cluster: [redacted] bgen_pdf

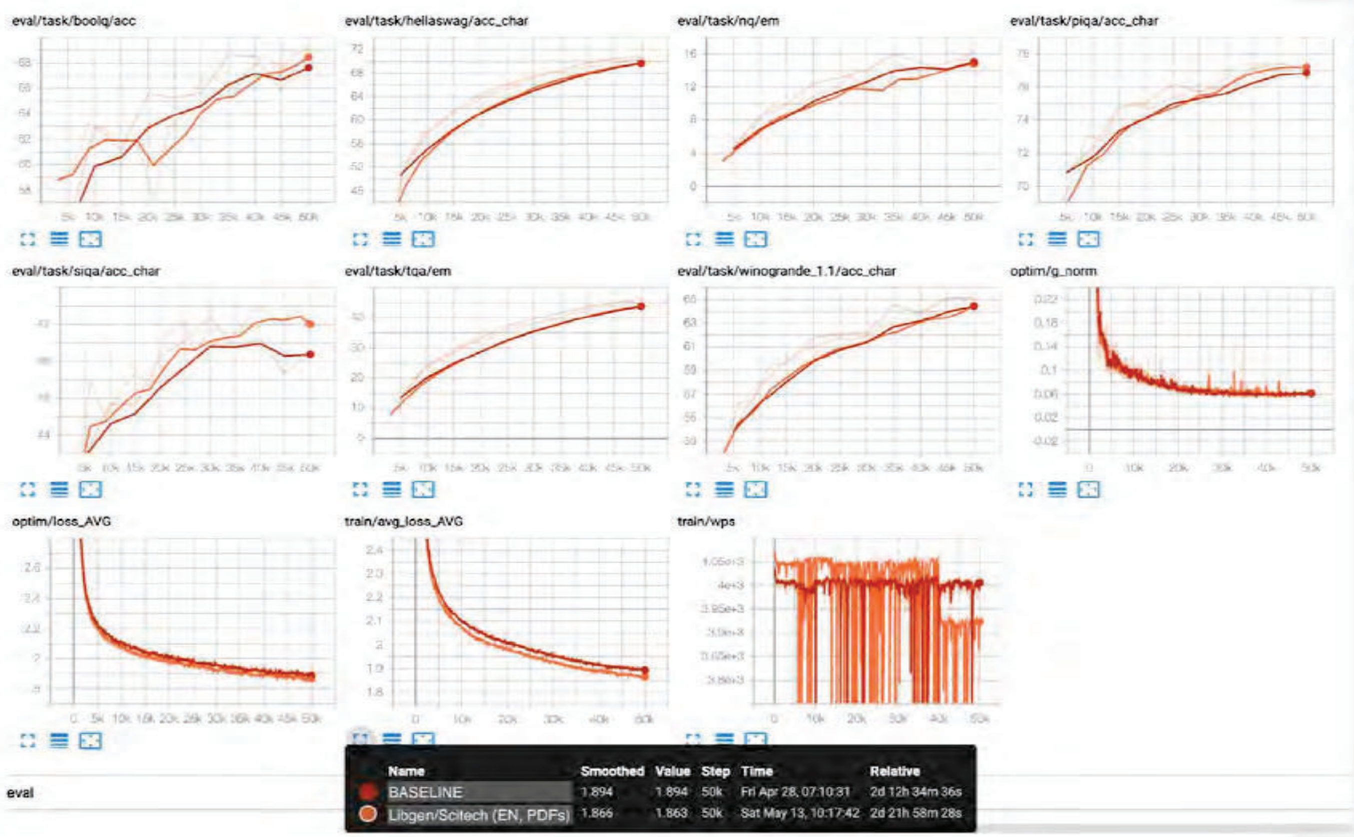
    - ○ 199,145 on [redacted] fair_llm [redacted]
- loaded previously EN PDF/EPUB on fair cluster (~480k): [redacted] ir_llm/data_v2/datasets/books

Fiction EN download status (5pm 15.05.2023): (33% done of 1,159,720)

- EPUBs (338,797) [redacted] /fiction/fiction_epub
- PDFs (44,109) : [redacted] iction/fiction_pdf

Ablation results for Scitech EN PDFs at 50k step (100% complete):

- Overall no red flags observed
- Some improvement on siqa and boolq (but that's within the stdev)
- TB: https://fburl.com/ [redacted]



Name	Smoothed	Value	Step	Time	Relative
● BASELINE	1.894	1.894	50k	Fri Apr 28, 07:10:31	2d 12h 34m 36s
○ Libgen/Scitech (EN, PDFs)	1.866	1.863	50k	Sat May 13, 10:17:42	2d 21h 58m 28s

12.05.2023

[Nikolay]
We have overall downloaded 1.6M books EN PDFs and EPUBs for Scitech (or 92%). This number however contains ~10% of corrupted file which needs to be re-downloaded later on (or skipped if they are corrupted in the source)
Scitech EN download status (12pm 12.05.2023): 92% done

- EPUBs (305k)
  - 111,272 on FAIR Cluster
  - 199,145 on RSC
- PDFs (810k)
  - 638,350 on Fair Cluster
  - 199,145 on RSC
- loaded previously EN PDF/EPUB on fair cluster (~480k): [redacted] fair_llm/data_v2/datasets/books

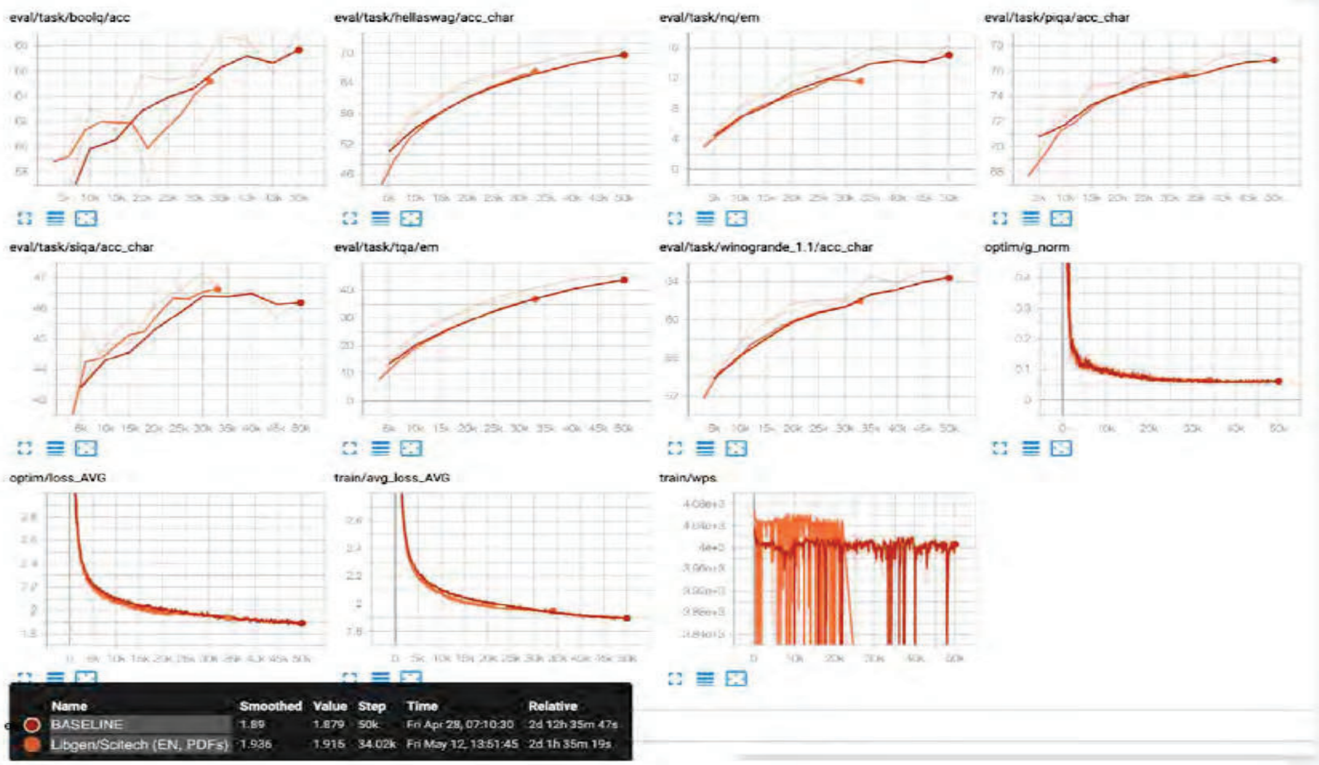
Ablation results at 35k step (70% complete):

- Overall no red flags observed
- Loss seems to flatten out earlier
- TB: https://fburl.com[redacted]

[Lukas]
Scitech EN PDF conversion status (2pm 12.05.2023):

- 350k books finished (52B tokens)
- 390k books ready to process

[Nikolay] ablation results:



	Name	Smoothed	Value	Step	Time	Relative
●	BASELINE	1.89	1.879	50k	Fri Apr 28, 07:10:30	2d 12h 35m 47s
●	Libgen/Scitech (EN, PDFs)	1.936	1.915	34.02k	Fri May 12, 13:51:45	2d 1h 35m 19s

11.05.2023

[Nikolay]

We will need to reload the corrupted files separately after going through the first round of parsing.
10% of files are corrupted files after initial download, both EPUBs and PDFs.
For EPUBs processing we used Marie-Anne's html2latex.py script (the one used for CC) and performed some post processing on top of it - removing the Copyright section.

Scitech download status (10pm 11.05.2023):

- EPUBs (305k)
  - 106,677 on FAIR Cluster
  - 199,145 on RSC
- PDFs (810k)
  - 611,409 on Fair Cluster
  - 199,145 on RSC
- loaded previously EN PDF/EPUB on fair cluster (~480k):[redacted]fair_llm[redacted]atasets/books

[Lukas]
PDFs:
Improved post-processing to remove all kinds of repeated patterns and more.

- 260k PDF books successfully parsed
- [redacted]atasets/books/data/scitech_pdf_ocr_all

[Peter]
EPUBs: prepared a script to postprocess the EPUBs.

- 82k EPUB books parsed
- [redacted]ibgen_epub_parsed

PDFs:

- [redacted]datasets/books/data/scitech_pdf_ocr_pet

10.05.2023

[Nikolay]

Libgen Scitech PDFs:

Starting an ablation experiment for 10% of scitech (parsed pdfs). We substitute 10% from CCNet with Libgen scitech dataset (matching it to the target datasets proportion: 2T Total vs 200B Libgen Scitech -> 10%).

Data	Total dataset size (billion tokens)	Baseline (weights/%)	Experiment (weights/%)	Epochs (# / 200B)
Stack Exchange	25	1.2 (1.8%)	1.2 (1.8%)	0.14
B3G (books3 + gutenberg)	28	3 (4.5%)	3 (4.5%)	0.3
Arxiv	33	1.6 (2.4%)	1.6 (2.4%)	0.15
Github OSS	271	3 (4.5%)	3 (4.5%)	0.03
C4 en	198	10 (15%)	10 (15%)	0.15
CCNet	1,416	45 (67%)	38 (57%)	0.08
Wikipedia	33	3 (4.5%)	3 (4.5%)	0.27

Libgen Scitech	25B (total: ~200B)	-	7 (10%)	0.8
Total	2.2T	67	100%	

Run (TB: https://fburl.com/[redacted]

- Libgen 10%: nb_7B_libgen_1005_run000
- Baseline: nb_7B_baseline

```
[redacted]fair_llm/xldumps/nb_7B_libgen_1005/[redacted]
pytho[redacted]gen_1005 train.py --sweep
[redacted]bgen_230510_7B_b4M_256gpu.yaml --mem 480 --ncpu 10 --ngpu 8 --ntasks 256 --nodes
32 --partition learn --anaconda[redacted]-qos fair_llm --launch_restart_dependencies
2
```

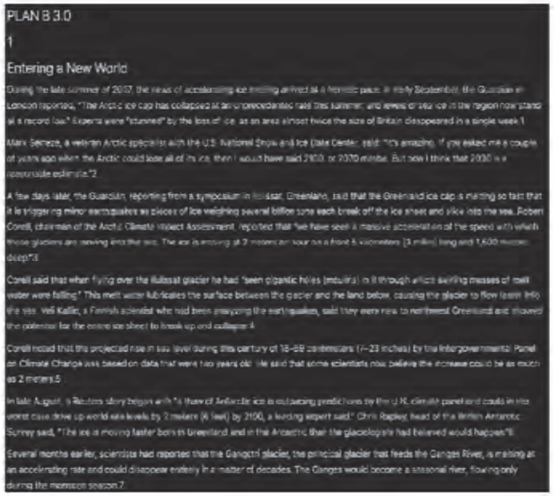
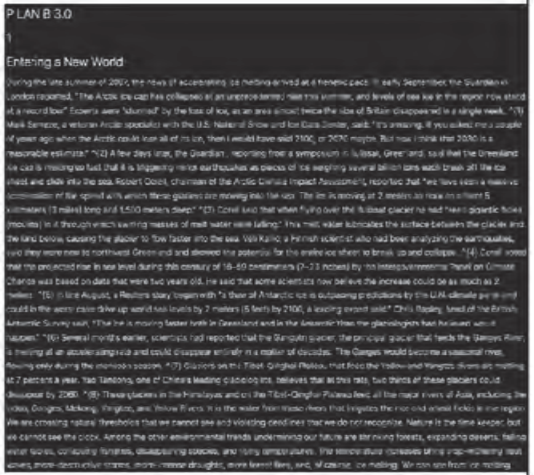
Dat[redacted]

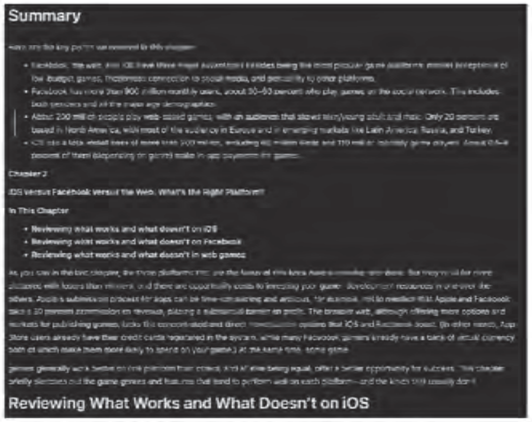
- [redacted]libgen_scitech/scitech_10_pct
- 106B chars in total -> 25B tokens

Config: https://www.internalfb.com/[redacted]

Libgen Scitech EPUBs:

The goal is to apply the same html_to_latex parser from CCNET.

html2text parsing (import html2text)	CC html_to_latex parsing (crawl.utils.html_to_latex)
 PLAN B 3.0 1 Entering a New World During the late summer of 2007, the news of accelerating ice melting arrived at a frenetic pace. In early September, the Guardian in London reported, "The Arctic ice cap has collapsed at an unprecedented rate this summer, and levels of sea ice in the region now stand at a record low." Experts were "stunned" by the loss of ice, as an area almost twice the size of Britain disappeared in a single week.1 Mark Serreze, a veteran Arctic specialist with the U.S. National Snow and Ice Data Center, said: "It's amazing. If you asked me a couple of years ago when the Arctic could lose all of its ice, then I would have said 2100, or 2070 maybe. But now I think that 2030 is a reasonable estimate."2 A few days later, the Guardian, reporting from a symposium in Ilulissat, Greenland, said that the Greenland ice cap is melting so fast that it is triggering minor earthquakes as pieces of ice weighing several billion tons each break off the ice sheet and slide into the sea. Robert Corell, chairman of the Arctic Climate Impact Assessment, reported that "we have seen a massive acceleration of the speed with which these glaciers are moving into the sea. The ice is moving at 2 meters an hour on a front 5 kilometers [3 miles] long and 1,500 meters deep."3 Corell said that when flying over the Ilulissat glacier he had "seen gigantic holes (moulins) in it through which swirling masses of melt water were falling." This melt water lubricates the surface between the glacier and the land below, causing the glacier to flow faster into the sea. Veli Kallio, a Finnish scientist who had been analyzing the earthquakes, said they were new to northwest Greenland and showed the potential for the entire ice sheet to break up and collapse.4 Corell noted that the projected rise in sea level during this century of 18–59 centimeters (7–23 inches) by the Intergovernmental Panel on Climate Change was based on data that were two years old. He said that some scientists now believe the increase could be as much as 2 meters.5 In late August, a Reuters story began with "A thaw of Antarctic ice is outpacing predictions by the U.N. climate panel and could in the worst case drive up world sea levels by 2 meters (6 feet) by 2100, a leading expert said." Chris Rapley, head of the British Antarctic Survey said, "The ice is moving faster both in Greenland and in the Antarctic than the glaciologists had believed would happen."6 Several months earlier, scientists had reported that the Gangotri glacier, the principal glacier that feeds the Ganges River, is melting at an accelerating rate and could disappear entirely in a matter of decades. The Ganges would become a seasonal river, flowing only during the monsoon season.7 074C98827769604025AAED5688612F4D	 P LAN B 3.0 1 Entering a New World During the late summer of 2007, the news of accelerating ice melting arrived at a frenetic pace. In early September, the Guardian in London reported, "The Arctic ice cap has collapsed at an unprecedented rate this summer, and levels of sea ice in the region now stand at a record low." Experts were "stunned" by the loss of ice, as an area almost twice the size of Britain disappeared in a single week. ^(1) Mark Serreze, a veteran Arctic specialist with the U.S. National Snow and Ice Data Center, said: "It's amazing. If you asked me a couple of years ago when the Arctic could lose all of its ice, then I would have said 2100, or 2070 maybe. But now I think that 2030 is a reasonable estimate." ^(2) A few days later, the Guardian , reporting from a symposium in Ilulissat, Greenland, said that the Greenland ice cap is melting so fast that it is triggering minor earthquakes as pieces of ice weighing several billion tons each break off the ice sheet and slide into the sea. Robert Corell, chairman of the Arctic Climate Impact Assessment, reported that "we have seen a massive acceleration of the speed with which these glaciers are moving into the sea. The ice is moving at 2 meters an hour on a front 5 kilometers (3 miles) long and 1,500 meters deep." ^(3) Corell said that when flying over the Ilulissat glacier he had "seen gigantic holes (moulins) in it through which swirling masses of melt water were falling." This melt water lubricates the surface between the glacier and the land below, causing the glacier to flow faster into the sea. Veli Kallio, a Finnish scientist who had been analyzing the earthquakes, said they were new to northwest Greenland and showed the potential for the entire ice sheet to break up and collapse. ^(4) Corell noted that the projected rise in sea level during this century of 18–59 centimeters (7–23 inches) by the Intergovernmental Panel on Climate Change was based on data that were two years old. He said that some scientists now believe the increase could be as much as 2 meters. ^(5) In late August, a Reuters story began with "a thaw of Antarctic ice is outpacing predictions by the U.N. climate panel and could in the worst case drive up world sea levels by 2 meters (6 feet) by 2100, a leading expert said." Chris Rapley, head of the British Antarctic Survey said, "The ice is moving faster both in Greenland and in the Antarctic than the glaciologists had believed would happen." ^(6) Several months earlier, scientists had reported that the Gangotri glacier, the principal glacier that feeds the Ganges River, is melting at an accelerating rate and could disappear entirely in a matter of decades. The Ganges would become a seasonal river, flowing only during the monsoon season. ^(7) Glaciers on the Tibet-Qinghai Plateau that feed the Yellow and Yangtze Rivers are melting at 7 percent a year. Yao Tandong, one of China's leading glaciologists, believes that at this rate, two thirds of these glaciers could disappear by 2060. ^(8) These glaciers in the Himalayas and on the Tibet-Qinghai Plateau feed all the major rivers of Asia, including the Indus, Ganges, Mekong, Yangtze, and Yellow Rivers. It is the water from these rivers that irrigates the rice and wheat fields in the region. We are crossing natural thresholds that we cannot see and violating deadlines that we do not recognize. Nature is the time keeper, but we cannot see the clock. Among the other environmental trends undermining our future are shrinking forests, expanding deserts, falling water tables, collapsing fisheries, disappearing species, and rising temperatures. The temperature increases bring crop-withering heat waves, more-destructive storms, more-intense droughts, more forest fires, and, of course, ice melting. We can see from ice melting



Summary

Here are the key points we covered in this chapter:

- Facebook, the web, and iOS have three major advantages besides being the most popular game platforms: market acceptance of low-budget games, seamless connection to social media, and portability to other platforms.
- Facebook has more than 900 million monthly users, about 30–60 percent who play games on the social network. This includes both genders and all the major age demographics.
- About 200 million people play web-based games, with an audience that skews teen/young adult and male. Only 20 percent are based in North America, with most of the audience in Europe and in emerging markets like Latin America, Russia, and Turkey.
- iOS has a total install base of more than 200 million, including 80 million iPads and 110 million monthly game players. About 0.5–6 percent of them (depending on genre) make in-app payments for games.

Chapter 2

iOS versus Facebook versus the Web: What's the Right Platform?

In This Chapter

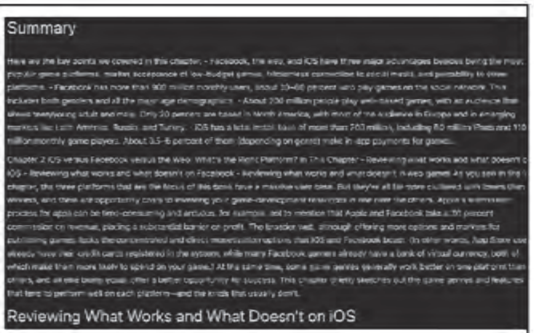
- **Reviewing what works and what doesn't on iOS**
- **Reviewing what works and what doesn't on Facebook**
- **Reviewing what works and what doesn't in web games**

As you saw in the last chapter, the three platforms that are the focus of this book have a massive user base. But they're all far more cluttered with losers than winners, and there are opportunity costs to investing your game-development resources in one over the others. Apple's submission process for apps can be time-consuming and arduous, for example, not to mention that Apple and Facebook take a 30 percent commission on revenue, placing a substantial barrier on profit. The broader web, although offering more options and markets for publishing games, lacks the concentrated and direct monetization options that iOS and Facebook boast. (In other words, App Store users already have their credit cards registered in the system, while many Facebook gamers already have a bank of virtual currency, both of which make them more likely to spend on your game.) At the same time, some game

genres generally work better on one platform than others, and all else being equal, offer a better opportunity for success. This chapter briefly sketches out the game genres and features that tend to perform well on each platform—and the kinds that usually don't.

Reviewing What Works and What Doesn't on iOS

0C7955E04686D9FD4CF2C2FA5D1B390C

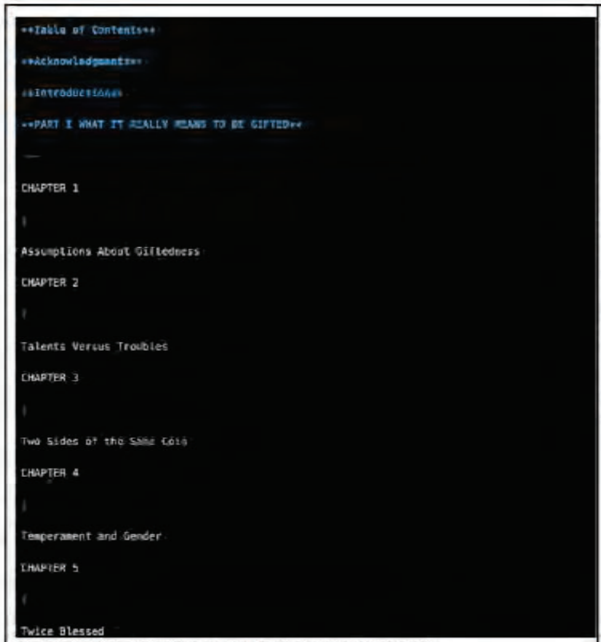


Summary

Here are the key points we covered in this chapter: - Facebook, the web, and iOS have three major advantages besides being the most popular game platforms: market acceptance of low-budget games, seamless connection to social media, and portability to other platforms. - Facebook has more than 900 million monthly users, about 30–60 percent who play games on the social network. This includes both genders and all the major age demographics. - About 200 million people play web-based games, with an audience that skews teen/young adult and male. Only 20 percent are based in North America, with most of the audience in Europe and in emerging markets like Latin America, Russia, and Turkey. - iOS has a total install base of more than 700 million, including 80 million iPads and 110 million monthly game players. About 3.5–6 percent of them (depending on genre) make in-app payments for games.

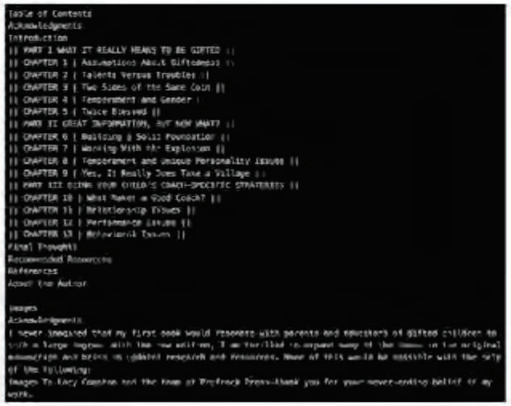
Chapter 2 iOS versus Facebook versus the Web: What's the Right Platform? In This Chapter - Reviewing what works and what doesn't o iOS - Reviewing what works and what doesn't on Facebook - Reviewing what works and what doesn't in web games As you saw in the chapter, the three platforms that are the focus of this book have a massive user base. But they're all far more cluttered with losers than winners, and there are opportunity costs to investing your game-development resources in one over the others. Apple's submission process for apps can be time-consuming and arduous, for example, not to mention that Apple and Facebook take a 30 percent commission on revenue, placing a substantial barrier on profit. The broader web, although offering more options and markets for publishing games, lacks the concentrated and direct monetization options that iOS and Facebook boast. (In other words, App Store use already have their credit cards registered in the system, while many Facebook gamers already have a bank of virtual currency, both of which make them more likely to spend on your game.) At the same time, some game genres generally work better on one platform than others, and all else being equal, offer a better opportunity for success. This chapter briefly sketches out the game genres and features that tend to perform well on each platform—and the kinds that usually don't.

Reviewing What Works and What Doesn't on iOS



```
**Table of Contents**

**Acknowledgments**

**Introduction**

**PART I WHAT IT REALLY MEANS TO BE GIFTED**

---

CHAPTER 1

|

Assumptions About Giftedness

CHAPTER 2

|

Talents Versus Troubles

CHAPTER 3

|

Two Sides of the Same Coin

CHAPTER 4

|

Temperament and Gender

CHAPTER 5

|

Twice Blessed
```

086018DED0E91FEB464DF69F059A92B7



```
Table of Contents
Acknowledgments
Introduction
|| PART I WHAT IT REALLY MEANS TO BE GIFTED ||
|| CHAPTER 1 | Assumptions About Giftedness ||
|| CHAPTER 2 | Talents Versus Troubles ||
|| CHAPTER 3 | Two Sides of the Same Coin ||
|| CHAPTER 4 | Temperament and Gender ||
|| CHAPTER 5 | Twice Blessed ||
|| PART II GREAT INFORMATION, BUT NOW WHAT? ||
|| CHAPTER 6 | Building a Solid Foundation ||
|| CHAPTER 7 | Working With the Explosion ||
|| CHAPTER 8 | Temperament and Unique Personality Issues ||
|| CHAPTER 9 | Yes, It Really Does Take a Village ||
|| PART III GIVING YOUR CHILD'S COACH-SPECIFIC STRATEGIES ||
|| CHAPTER 10 | What Makes a Good Coach? ||
|| CHAPTER 11 | Relationship Issues ||
|| CHAPTER 12 | Performance Issues ||
|| CHAPTER 13 | Behavioral Issues ||
Final Thoughts
Recommended Resources
References
About the Author

Images
Acknowledgments
I never imagined that my first book would resonate with parents and educators of gifted children to
such a large degree. With the new edition, I am thrilled to expand many of the ideas in the original
manuscript and bring in updated research and resources. None of this would be possible with the help
of the following:
Images To Lacy Compton and the team at Prufrock Press—thank you for your never-ending belief in my
work.
```

09.05.2023

[Nikolay]

Decided to go with the direct file upload without using torrents for the following reasons:

- using torrents would entail "seeding" the files - i.e. sharing the content outside, this could be legally not OK
- with the direct file download we can pre-filter the needed format and language of the files - i.e. downloading only EN, PDF and EPUB initially
- the downside is that this way it is slower and need more engineering to bypass IP throttling and download retries
- we can reload specific MD5 file names, that were corrupted or missing from the initial download (based on Lukas's observations there are 30% of corrupted files in the initial Libgen download)

Currently loading using 2 dev machines and 1 fair cluster. Approximately an additional 10TB of data loaded (1M books out of 1.3M): 75% of EN, PDF or EPUB scitech books.

Raw downloaded data locations:

- ~800k EN books on fair cluster [redacted]
    - 546k pdfs
    - 80k epubs
    - 8k corrupted files
- ~400k EN books on [redacted]
    - 200k pdfs
    - 200k epubs
- loaded previously EN PDF/EPUB on fair cluster (~480k): [redacted]latasets/books

Parsed data:

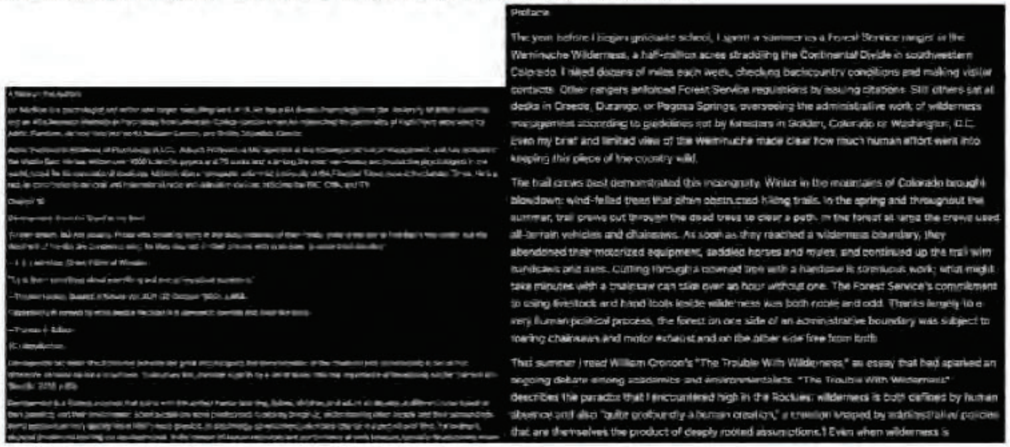
- 10% scitech (pdfs only): [redacted]ibgen/scitech_10_pct/

Total numbers (in # of books):

- Libgen: 3.7M

- Libgen (EN & PDF/EPUB): 1.7M | Downloaded 1.3M
- Libgen (EN & PDF): 1.3M -> parsed 13%

Examples of parsed EPUBs (light version of parsing w/o M-A's script):



[illegible]

Preface

The year before I began graduate school, I spent a summer as a Forest Service ranger in the Weminuche Wilderness, a half-million acres straddling the Continental Divide in southwestern Colorado. I hiked dozens of miles each week, checking backcountry conditions and making visitor contacts. Other rangers enforced Forest Service regulations by issuing citations. Still others sat at desks in Creede, Durango, or Pagosa Springs, overseeing the administrative work of wilderness management according to guidelines set by foresters in Golden, Colorado or Washington, D.C. Even my brief and limited view of the Weminuche made clear how much human effort went into keeping this piece of the country wild.

The trail crews best demonstrated this incongruity. Winter in the mountains of Colorado brought blowdown: wind-felled trees that often obstructed hiking trails. In the spring and throughout the summer, trail crews cut through the dead trees to clear a path. In the forest at large the crews used all-terrain vehicles and chainsaws. As soon as they reached a wilderness boundary, they abandoned their motorized equipment, saddled horses and mules, and continued up the trail with handsaws and axes. Cutting through a downed tree with a handsaw is strenuous work; what might take minutes with a chainsaw can take over an hour without one. The Forest Service's commitment to using livestock and hand tools inside wilderness was both noble and odd. Thanks largely to a very human political process, the forest on one side of an administrative boundary was subject to roaring chainsaws and motor exhaust and on the other side free from both.

That summer I read William Cronon's "The Trouble With Wilderness," an essay that had sparked an ongoing debate among academics and environmentalists. "The Trouble With Wilderness" describes the paradox that I encountered high in the Rockies: wilderness is both defined by human absence and also "quite profoundly a human creation," a creation shaped by administrative policies that are themselves the product of deeply rooted assumptions.1 Even when wilderness is

[Lukas]
Filtered scitech conversion is 75% done out of the first chunk of 340k EN PDF books (total chunk size of scitech EN PDFs: 1.3M, so we've parsed ~13% of EN PDFs). We pre-selected 340k books (PDF). 34% of the files are corrupted. Finished 167k (uncorrupted) books.

Conversion speed:

- Ideal: 6.8 ± 1.9 PDF / GPU*h
- Actual (b/c of insufficient number of GPUs): 2.2 PDF / GPU*h

Commented [6]: That paints a wrong picture. The speed per GPU is still around 7 books per hour. The number of GPUs is the bottleneck

Commented [7]: fair poin [redacted] a ballpark how much more we need [redacted]@meta.com was mentioning that we c PUs from the Retina team

Commented [8]: As a ball park estimate:
we would need ~2k GPUs on FAIR Cluster for 2 weeks starting from ~mid-next week.
It's actually similar to what we use now for sci-tech, so we might keep the current strategy of just asking [redacted] run from their accounts.
[redacted]@meta.com [redacted]@meta.com

Implemented an additional step of post processing to remove repeated reference items.
Combined directory: scitech_pdf_ocr_all

Processed chunks (~10% of scitech):

- [redacted]books/data/scitech_pdf_ocr_jsonl/chunks

05.05.2023

Launched slurm jobs for OCR parsing of the first 15% of Libgen:

- scitech_pdf_ocr: first half of parsed files
- scitech_pdf_ocr_af: second half of parsed files

Slurm job command

04.05.2023

Plan:

- [Nikolay] check about gpus on fair cluster -> how much we can use: 1k GPUs - DONE
- [Lukas] prepare 3 sample pages of books with formulas, tables and lists original VS parsed with small OCR model - DONE
- [Nikolay] Pre-filter data to only EN (since OCR parsing works best with EN) -> We will pre-filter EN and PDFs only as the OCR script works best with EN. - DONE
- [Lukas][Nikolay] start the pipeline for parsing first 10% of PDFs on fair cluster: use the current dump of PDFs: [redacted]datasets/books - DONE
- [Lukas] add the token to split the sequence (in case the page was skipped due to parsing error) - DONE
- [Nikolay] prepare pipeline for loading remaining data from libgen DONE
- [Nikolay] prepare pipeline for parsing EPUBs
- [Nikolay] -> run ablations for processed PDFs

Fastext classifier for language:

Weights (on FAIR cluster)[redacted]air_llm/datasets/tools
GitHub Fasttext code:
git clone https://github.com/facebookresearch[redacted]
cd fastText
make
pip install .

Observed OCR parsing artifacts:

- 0ab5ee32e73a2a455e0cc14894462f69.pdf: In References all references are duplicated

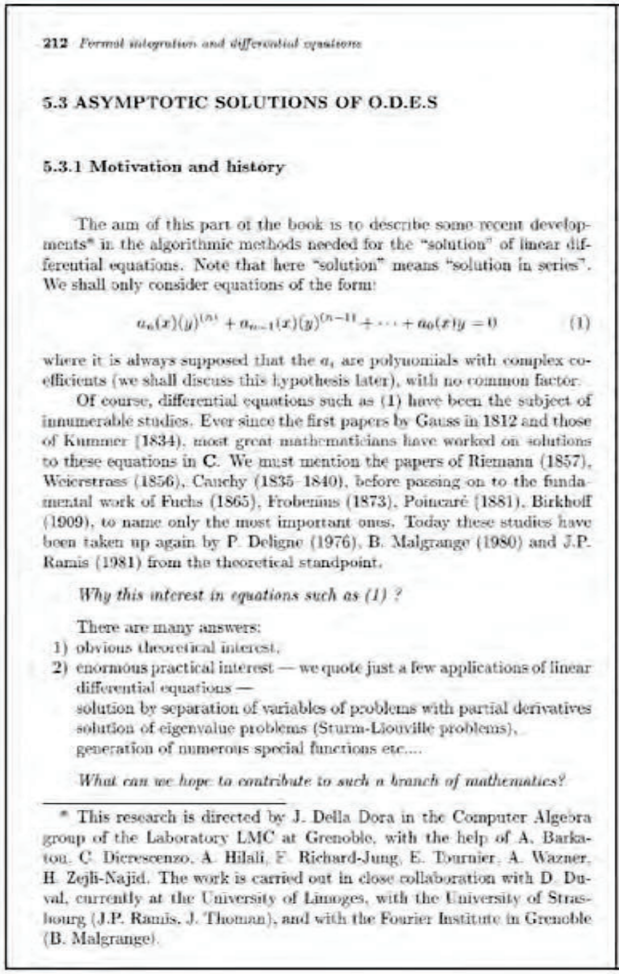
[Nikolay] Loaded 3% of the sci-tech libgen library:

- PDFs: 600GB, 66332 files
- EPUBs: 1.5GB, 781

[Lukas] Smaller model metrics are on par with base model now. Retrained with larger training set.
Model speed ~1.8k pages / gpu*hour → 2.2x speed up
[redacted]dataset/scitech/mmd_small2
per page - i.e. pages are not parsed and skipped

Examples of OCR Parsing
Random books from scitech, pages chosen for diversity

ORIGINAL	PARSED WITH OCR



212 *Formal integration and differential equations*

5.3 ASYMPTOTIC SOLUTIONS OF O.D.E.S

5.3.1 Motivation and history

The aim of this part of the book is to describe some recent developments* in the algorithmic methods needed for the "solution" of linear differential equations. Note that here "solution" means "solution in series". We shall only consider equations of the form:

$$a_n(x)(y)^{(n)} + a_{n-1}(x)(y)^{(n-1)} + \cdots + a_0(x)y = 0 \qquad (1)$$

where it is always supposed that the a_i are polynomials with complex coefficients (we shall discuss this hypothesis later), with no common factor.

Of course, differential equations such as (1) have been the subject of innumerable studies. Ever since the first papers by Gauss in 1812 and those of Kummer (1834), most great mathematicians have worked on solutions to these equations in **C**. We must mention the papers of Riemann (1857), Weierstrass (1856), Cauchy (1835–1840), before passing on to the fundamental work of Fuchs (1865), Frobenius (1873), Poincaré (1881), Birkhoff (1909), to name only the most important ones. Today these studies have been taken up again by P. Deligne (1976), B. Malgrange (1980) and J.P. Ramis (1981) from the theoretical standpoint.

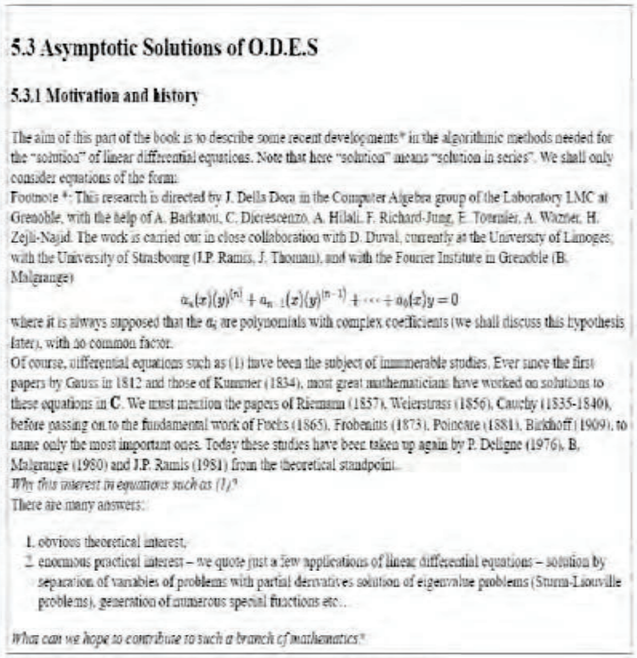
Why this interest in equations such as (1) ?

There are many answers:

1) obvious theoretical interest,
2) enormous practical interest — we quote just a few applications of linear differential equations —
solution by separation of variables of problems with partial derivatives
solution of eigenvalue problems (Sturm-Liouville problems),
generation of numerous special functions etc....

What can we hope to contribute to such a branch of mathematics?

* This research is directed by J. Della Dora in the Computer Algebra group of the Laboratory LMC at Grenoble, with the help of A. Barkatou, C. Dicrescenzo, A. Hilali, F. Richard-Jung, E. Tournier, A. Wazner, H. Zejli-Najid. The work is carried out in close collaboration with D. Duval, currently at the University of Limoges, with the University of Strasbourg (J.P. Ramis, J. Thomann), and with the Fourier Institute in Grenoble (B. Malgrange).



5.3 Asymptotic Solutions of O.D.E.S

5.3.1 Motivation and history

The aim of this part of the book is to describe some recent developments* in the algorithmic methods needed for the "solution" of linear differential equations. Note that here "solution" means "solution in series". We shall only consider equations of the form:
Footnote *: This research is directed by J. Della Dora in the Computer Algebra group of the Laboratory LMC at Grenoble, with the help of A. Barkatou, C. Dicrescenzo, A. Hilali, F. Richard-Jung, E. Tournier, A. Wazner, H. Zejli-Najid. The work is carried out in close collaboration with D. Duval, currently at the University of Limoges, with the University of Strasbourg (J.P. Ramis, J. Thomann), and with the Fourier Institute in Grenoble (B. Malgrange)

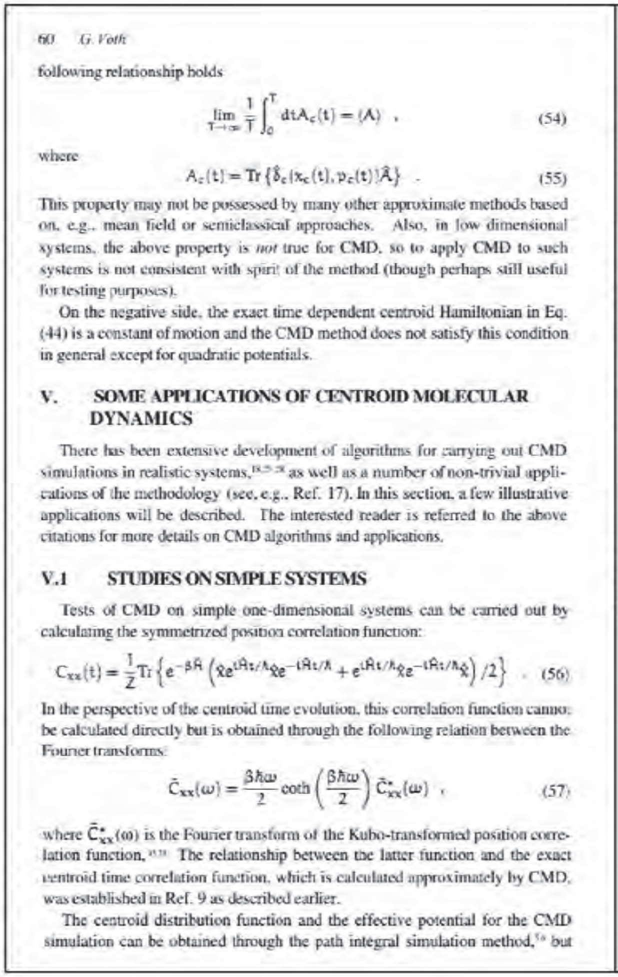
$$a_n(x)(y)^{(n)} + a_{n-1}(x)(y)^{(n-1)} + \cdots + a_0(x)y = 0$$

where it is always supposed that the a_i are polynomials with complex coefficients (we shall discuss this hypothesis later), with no common factor.
Of course, differential equations such as (1) have been the subject of innumerable studies. Ever since the first papers by Gauss in 1812 and those of Kummer (1834), most great mathematicians have worked on solutions to these equations in **C**. We must mention the papers of Riemann (1857), Weierstrass (1856), Cauchy (1835-1840), before passing on to the fundamental work of Fuchs (1865), Frobenius (1873), Poincare (1881), Birkhoff (1909), to name only the most important ones. Today these studies have been taken up again by P. Deligne (1976), B. Malgrange (1980) and J.P. Ramis (1981) from the theoretical standpoint.
Why this interest in equations such as (1)?
There are many answers:

1. obvious theoretical interest,
2. enormous practical interest – we quote just a few applications of linear differential equations – solution by separation of variables of problems with partial derivatives solution of eigenvalue problems (Sturm-Liouville problems), generation of numerous special functions etc...

What can we hope to contribute to such a branch of mathematics?

Note: Footnotes are placed after the paragraph



60 *G. Voth*

following relationship holds

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T dt A_c(t) = \langle A\rangle \quad , \tag{54}$$

where

$$A_c(t) = \mathrm{Tr}\{\hat\delta_c(x_c(t), p_c(t))\hat A\} \quad . \tag{55}$$

This property may not be possessed by many other approximate methods based on, e.g., mean field or semiclassical approaches. Also, in low dimensional systems, the above property is *not* true for CMD, so to apply CMD to such systems is not consistent with spirit of the method (though perhaps still useful for testing purposes).

On the negative side, the exact time dependent centroid Hamiltonian in Eq. (44) is a constant of motion and the CMD method does not satisfy this condition in general except for quadratic potentials.

V. SOME APPLICATIONS OF CENTROID MOLECULAR DYNAMICS

There has been extensive development of algorithms for carrying out CMD simulations in realistic systems,[18,27,28] as well as a number of non-trivial applications of the methodology (see, e.g., Ref. 17). In this section, a few illustrative applications will be described. The interested reader is referred to the above citations for more details on CMD algorithms and applications.

V.1 STUDIES ON SIMPLE SYSTEMS

Tests of CMD on simple one-dimensional systems can be carried out by calculating the symmetrized position correlation function:

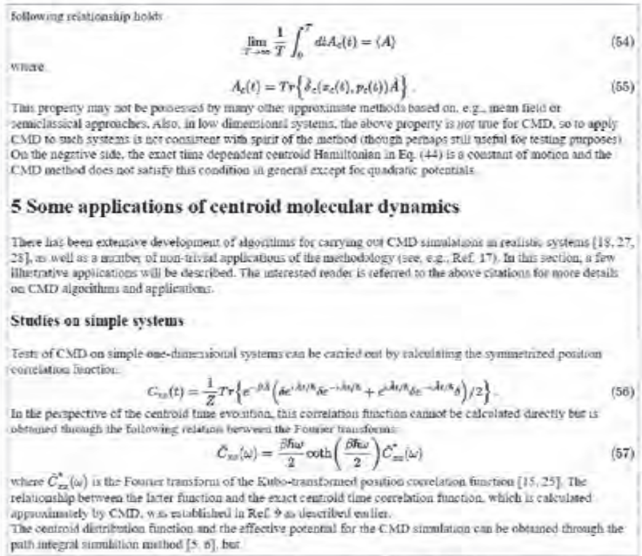
$$C_{xx}(t) = \frac{1}{Z}\mathrm{Tr}\left\{e^{-\beta\hat H}\left(\hat x e^{i\hat H t/\hbar}\hat x e^{-i\hat H t/\hbar} + e^{i\hat H t/\hbar}\hat x e^{-i\hat H t/\hbar}\hat x\right)/2\right\} \quad . \tag{56}$$

In the perspective of the centroid time evolution, this correlation function cannot be calculated directly but is obtained through the following relation between the Fourier transforms:

$$\tilde C_{xx}(\omega) = \frac{\beta\hbar\omega}{2}\coth\left(\frac{\beta\hbar\omega}{2}\right)\tilde C^{*}_{xx}(\omega) \quad , \tag{57}$$

where $\tilde C^{*}_{xx}(\omega)$ is the Fourier transform of the Kubo-transformed position correlation function.[15,25] The relationship between the latter function and the exact centroid time correlation function, which is calculated approximately by CMD, was established in Ref. 9 as described earlier.

The centroid distribution function and the effective potential for the CMD simulation can be obtained through the path integral simulation method,[5,6] but



following relationship holds

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T dt A_c(t) = \langle A\rangle \tag{54}$$

where

$$A_c(t) = Tr\left\{\hat\delta_c(x_c(t), p_c(t))\hat A\right\} . \tag{55}$$

This property may not be possessed by many other approximate methods based on, e.g., mean field or semiclassical approaches. Also, in low dimensional systems, the above property is *not* true for CMD, so to apply CMD to such systems is not consistent with spirit of the method (though perhaps still useful for testing purposes). On the negative side, the exact time dependent centroid Hamiltonian in Eq. (44) is a constant of motion and the CMD method does not satisfy this condition in general except for quadratic potentials.

5 Some applications of centroid molecular dynamics

There has been extensive development of algorithms for carrying out CMD simulations in realistic systems [18, 27, 28], as well as a number of non-trivial applications of the methodology (see, e.g., Ref. 17). In this section, a few illustrative applications will be described. The interested reader is referred to the above citations for more details on CMD algorithms and applications.

Studies on simple systems

Tests of CMD on simple one-dimensional systems can be carried out by calculating the symmetrized position correlation function:

$$C_{xx}(t) = \frac{1}{Z}Tr\left\{e^{-\beta\hat H}\left(\hat x e^{i\hat H t/\hbar}\hat x e^{-i\hat H t/\hbar} + e^{i\hat H t/\hbar}\hat x e^{-i\hat H t/\hbar}\hat x\right)/2\right\} . \tag{56}$$

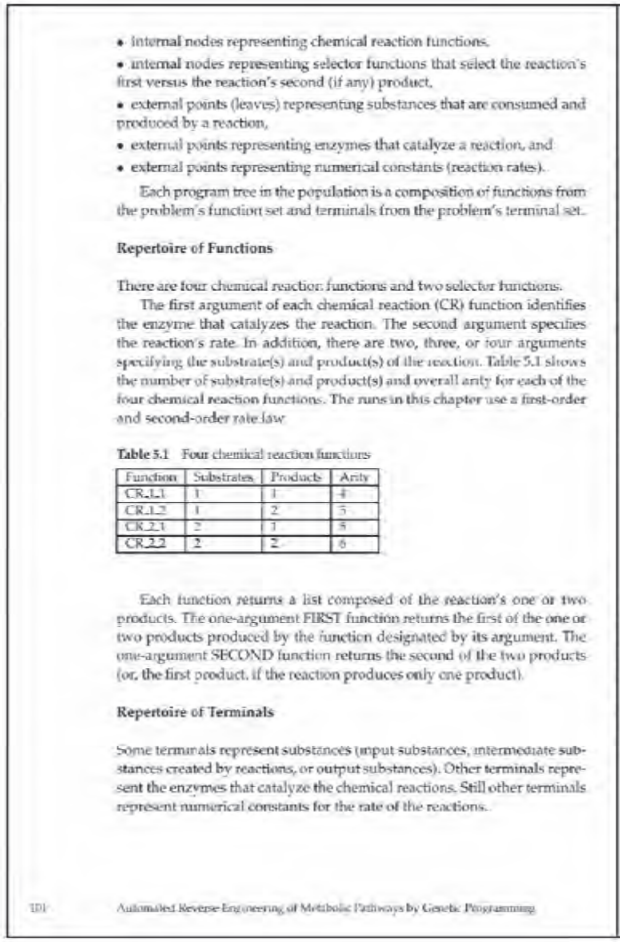
In the perspective of the centroid time evolution, this correlation function cannot be calculated directly but is obtained through the following relation between the Fourier transforms:

$$\tilde C_{xx}(\omega) = \frac{\beta\hbar\omega}{2}\coth\left(\frac{\beta\hbar\omega}{2}\right)\tilde C^{*}_{xx}(\omega) \tag{57}$$

where $\tilde C^{*}_{xx}(\omega)$ is the Fourier transform of the Kubo-transformed position correlation function [15, 25]. The relationship between the latter function and the exact centroid time correlation function, which is calculated approximately by CMD, was established in Ref. 9 as described earlier.

The centroid distribution function and the effective potential for the CMD simulation can be obtained through the path integral simulation method [5, 6], but

Note: In some cases the equation number is added, but not always. We can choose to remove all equation tags.



- internal nodes representing chemical reaction functions.
- internal nodes representing selector functions that select the reaction's first versus the reaction's second (if any) product,
- external points (leaves) representing substances that are consumed and produced by a reaction,
- external points representing enzymes that catalyze a reaction, and
- external points representing numerical constants (reaction rates).

Each program tree in the population is a composition of functions from the problem's function set and terminals from the problem's terminal set.

Repertoire of Functions

There are four chemical reaction functions and two selector functions.

The first argument of each chemical reaction (CR) function identifies the enzyme that catalyzes the reaction. The second argument specifies the reaction's rate. In addition, there are two, three, or four arguments specifying the substrate(s) and product(s) of the reaction. Table 5.1 shows the number of substrate(s) and product(s) and overall arity for each of the four chemical reaction functions. The runs in this chapter use a first-order and second-order rate law

Table 5.1 Four chemical reaction functions

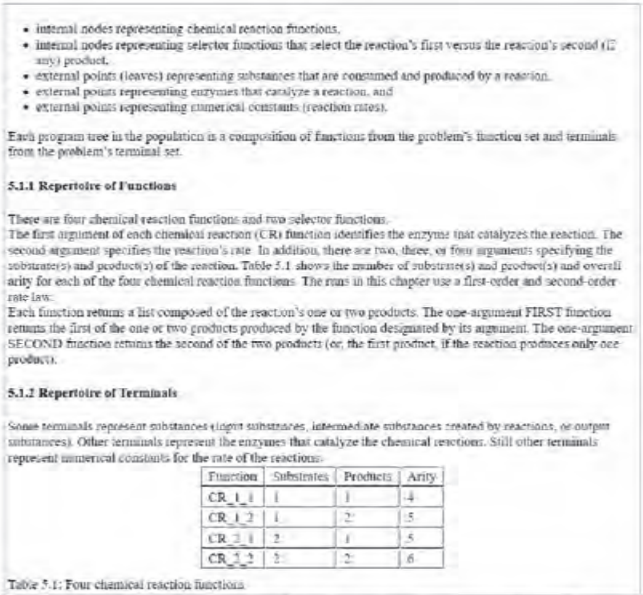
Function	Substrates	Products	Arity
CR.1.1	1	1	4
CR.1.2	1	2	5
CR.2.1	2	1	5
CR.2.2	2	2	6

Each function returns a list composed of the reaction's one or two products. The one-argument FIRST function returns the first of the one or two products produced by the function designated by its argument. The one-argument SECOND function returns the second of the two products (or, the first product, if the reaction produces only one product).

Repertoire of Terminals

Some terminals represent substances (input substances, intermediate substances created by reactions, or output substances). Other terminals represent the enzymes that catalyze the chemical reactions. Still other terminals represent numerical constants for the rate of the reactions.

101 Automated Reverse Engineering of Metabolic Pathways by Genetic Programming



- internal nodes representing chemical reaction functions.
- internal nodes representing selector functions that select the reaction's first versus the reaction's second (if any) product.
- external points (leaves) representing substances that are consumed and produced by a reaction.
- external points representing enzymes that catalyze a reaction, and
- external points representing numerical constants (reaction rates).

Each program tree in the population is a composition of functions from the problem's function set and terminals from the problem's terminal set.

5.1.1 Repertoire of Functions

There are four chemical reaction functions and two selector functions.
The first argument of each chemical reaction (CR) function identifies the enzyme that catalyzes the reaction. The second argument specifies the reaction's rate. In addition, there are two, three, or four arguments specifying the substrate(s) and product(s) of the reaction. Table 5.1 shows the number of substrate(s) and product(s) and overall arity for each of the four chemical reaction functions. The runs in this chapter use a first-order and second-order rate law.
Each function returns a list composed of the reaction's one or two products. The one-argument FIRST function returns the first of the one or two products produced by the function designated by its argument. The one-argument SECOND function returns the second of the two products (or, the first product, if the reaction produces only one product).

5.1.2 Repertoire of Terminals

Some terminals represent substances (input substances, intermediate substances created by reactions, or output substances). Other terminals represent the enzymes that catalyze the chemical reactions. Still other terminals represent numerical constants for the rate of the reactions.

Function	Substrates	Products	Arity
CR_1_1	1	1	4
CR_1_2	1	2	5
CR_2_1	2	1	5
CR_2_2	2	2	6

Table 5.1: Four chemical reaction functions

Note: Sometimes the model hallucinates subsection numbers (here from the table label) due to training data impurity. We can choose to filter out all section numbering.
Also, tables and figure captions will always be placed at the end of the page

02.05.2023

[Lukas] Smaller decoder model has a 2x greater conversion speed. Metrics are slightly worse but parsing samples look similar
PDF parsing samples smaller model[redacted]dataset/scitech/mmd_small

28.04.2023

[Lukas] Parsed with OCR library 70 books (29,488 pages total), it took 18 hours on 2 GPUs -> 2 books / gpu*hour -> ~800 pages / gpu*hour

- sci-tech: 3,274,071 books * 51% EN * 65% PDFs = 1M books = 260M pages
  260M pages / (500 pages / hour*gpu) = 500k GPU*[redacted] with 1000 GPUs it will take 500 hours (20 days)
  $25 / GPU day -> 1000*20*$25 = $0.5M (VS $16M
- sci-mag: 72,624,976 articles * 50% EN * 6 pages =
  220M pages / (500 pages / hour*gpu) = 440k GPU*hours -> 18 days with 1000 GPUs

PDF parsing samples: [redacted]dataset/scitech/mmd

26.04.2023

There is a sample of downloaded libgen documents on fair cluster (totals taken from here):
[redacted]data_v2/datasets/books
- fiction: 126GB (2% of total 5.6TB)

- scitech: 9.3TB (16% of total 59.4TB)
- scimag: 397GB (0.5% of total 80.6TB)

Fair cluster -> Python Lib torrent (list of magnet links) 50 torrents -> 2 days

Some processed samples from scitech on fair cluster:
[redacted]/data_v2/datasets/books/data/scitech_pdf/

- scitech processed PDFs: 63GB

24.04.2023

Reading metadata from the MySQL dumps: http://libgen.rs/dbdumps/. There are 3 category of content:

- Fiction: fiction.rar ->1,607,593 unique records (title&author)
- Scitech: libgen.rar -> 3,274,071 unique records (title&author)
- Scimag: scimag.sql.gz -> TBD

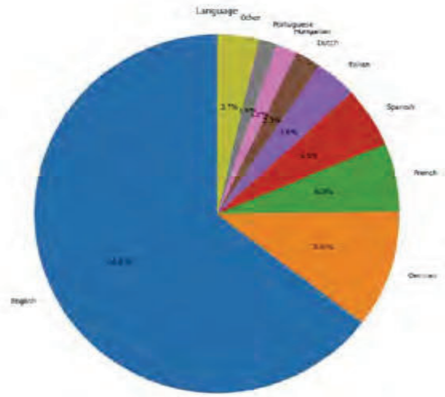
Findings:

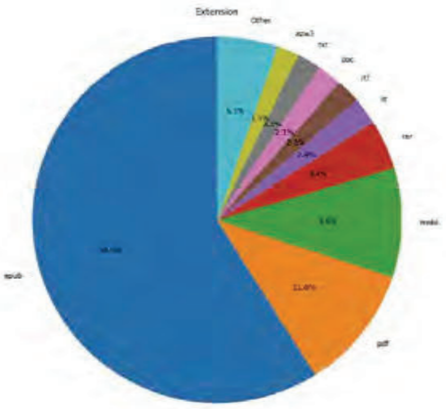
- Each DB dump contains metadata (table: fiction), book description (table: fiction_description) and hashes (table: fiction_hashes)
-
- Hashes table (fiction_hashes) provides the hashes to download files using torrents or IPFS (InterPlanetary File System -file sharing peer-to-peer network):
    - Torrent (using BitTorrent Info Hash: 'btih'): magnet:?xt=urn:btih:YOUR_BT_HASH -> paste this link into qBittorrent or µTorrent, or Transmission.
    - IPFS downloads (using 'ipfs_cid'): https://ipfs.io/ipfs/YOUR_IPFS_CID
    - Other columns: 'md5', 'crc32', 'edonkey', 'aich', 'sha1', 'tth', 'btih', 'sha256', 'ipfs_cid'

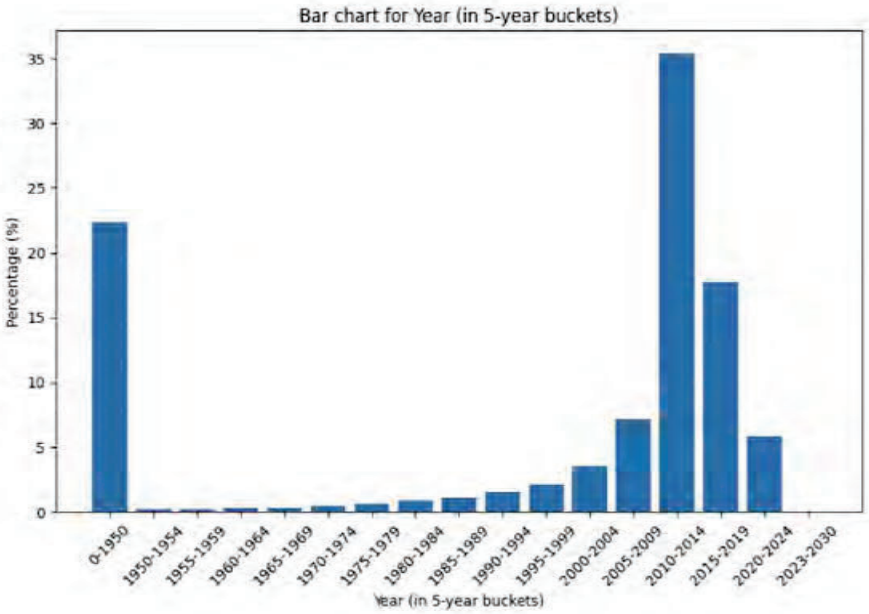
- LibGen is a different project and database from Sci-Hub. The sci-tech section of LibGen focuses on scientific and technical books, while the sci-mag section provides access to scientific and academic journal articles, which is the primary focus of Sci-Hub.

Fiction

- Tables: fiction, fiction_description, fiction_hashes
- fiction table num_records: 2,693,056
- columns: ['ID', 'MD5', 'Title', 'Author', 'Series', 'Edition', 'Language', 'Year', 'Publisher', 'Identifier', 'GooglebookID', 'ASIN', 'Coverurl', 'Extension', 'Filesize', 'Library', 'Issue', 'Locator', 'Commentary', 'Generic', 'Visible', 'TimeAdded', 'TimeLastModified']
- English: 65% | German: 11% | French: 6%
- Epub: 59% | PDF: 11% | mobi: 10%
- 0.5M books without a year


Language
Other
Portuguese
Hungarian
Dutch
Italian
Spanish
French
German
English
3.7%
2.3%

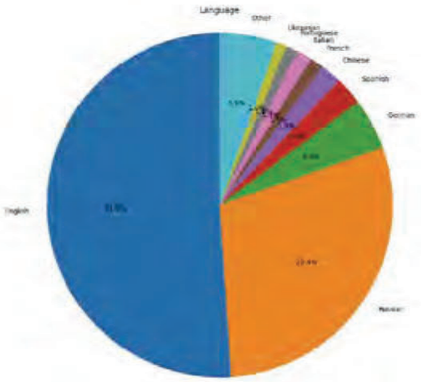

Extension
Other
azw3
txt
doc
rtf
lit
rar
mobi
pdf
epub
11.0%
2.1%
2.4%

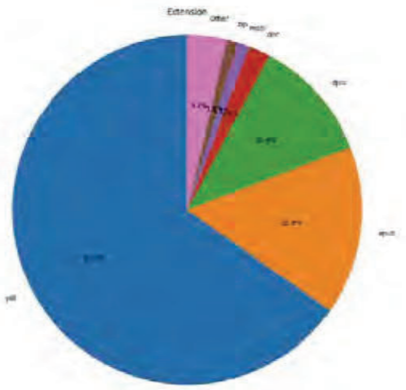


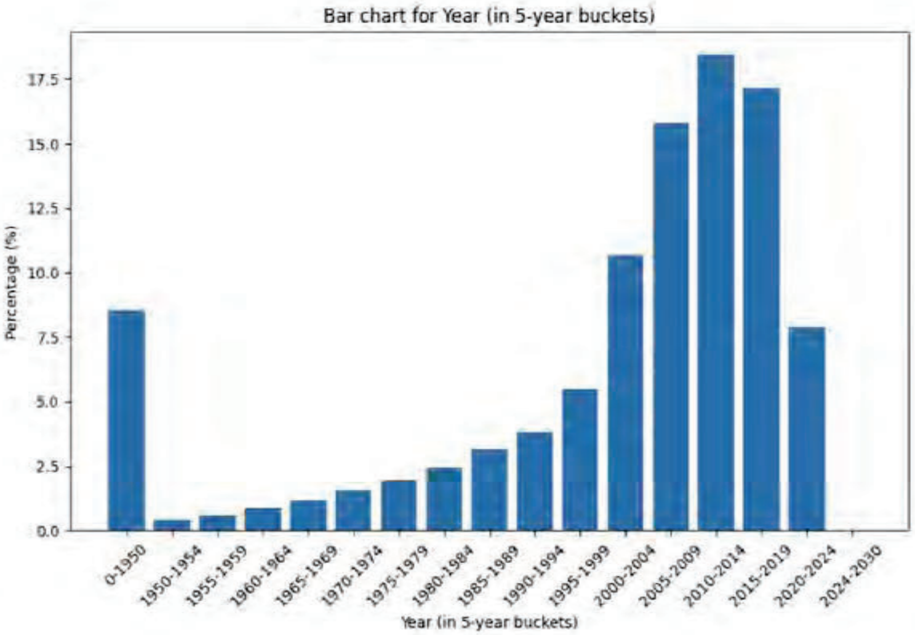
Sci-tech (libgen - main sci-tech collection)

Description: https://wiki.mhut.org/catalog:database

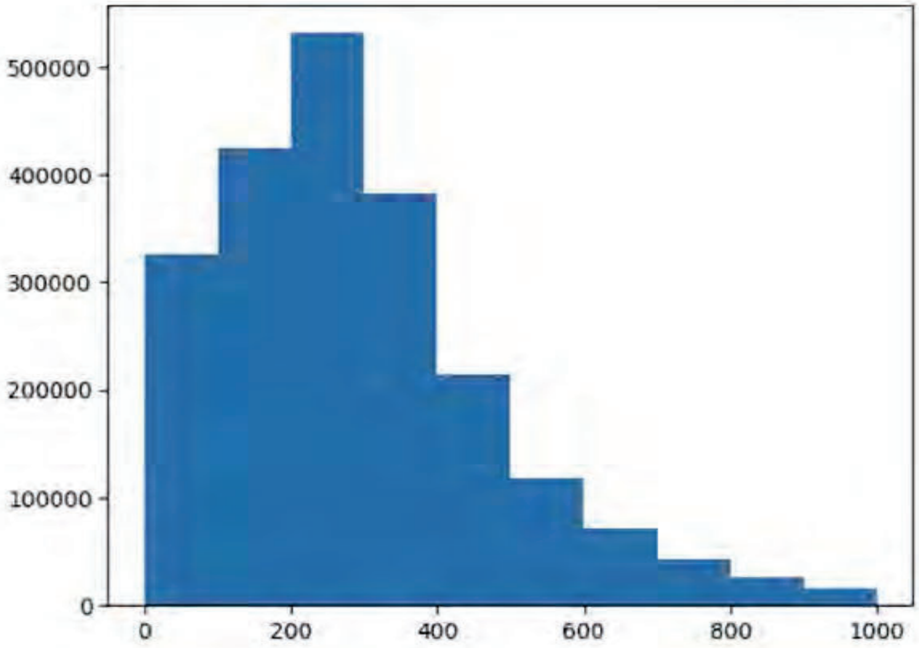
- Tables: updated (main metadata table), updated_edited, description, description_edited, hashes, topics
- updated table num_records: 3,706,772
- English: 51% | Russian 29% | German: 5%
- Epub: 16% | PDF: 65% | djvu: 11%


Language
Other
Ukrainian
Portuguese
Italian
French
Chinese
Spanish
German
Russian
English


Extension
Other
zip
mobi
pdf



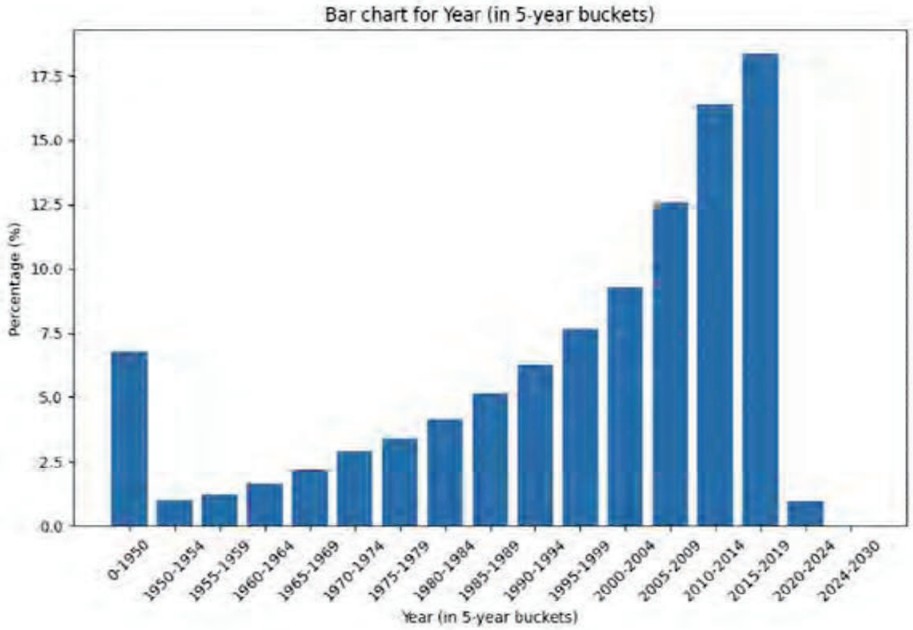
Pages distribution



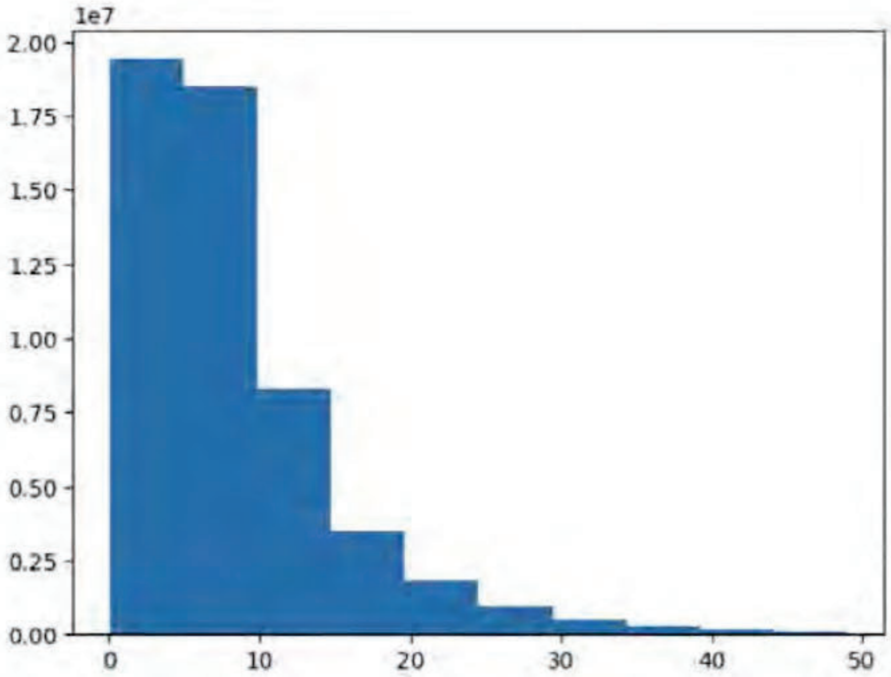
Sci-mag

- Tables: scimag, publishers, magazines, error_report
- fiction table num_records: 2,693,056

- columns: ['ID', 'MD5', 'Title', 'Author', 'Series', 'Edition', 'Language', 'Year', 'Publisher', 'Identifier', 'GooglebookID', 'ASIN', 'Coverurl', 'Extension', 'Filesize', 'Library', 'Issue', 'Locator', 'Commentary', 'Generic', 'Visible', 'TimeAdded', 'TimeLastModified']
- English: 65% | German: 11% | French: 6%
- Epub: 59% | PDF: 11% | mobi: 10%
- Scientific articles in this dump are before May 2020



Pages distribution



21.04.2023

Key takeaways:

- Only 3% of books from [redacted] epub format are in LibGen (out of 1000 sample).

Data:

- http://libgen.rs/dbdumps/: libgen metadata dumps [loading this 1.1 GB fiction.rar file takes 10 hours □ - could use Folx to download in multiple threads]
- http://libgen.rs/scimag/repository_torrent/: torrent files for scimag
- https://phillm.net: some indexer of torrent seeds
- https://ipfs.io/ipfs/bafkreibjbw2czkimwt5q7yeu3wko3a2fuw6q4km7rwo2wweirc6oejmokm: candidat for metadata DB dump

19.04.2023

We need to come up with a reliable book matching algorithm. There are many books with similar titles (ex. C/C++), so we need to account for authors' matches as well (at least partial authors match). The matching algorithm used checks for the exact Title match and at least one of the authors match.

Results:

- Up to 90% of books are present in LibGen for [redacted] and up to 76% fo [redacted]
- The books in LibGen are in djvu/epub/pdf format, so the parsing quality would be worse compared to getting the books from publishers directly. However epub is almost the same as HTML - it's a ZIP archive containing a collection of HTML, CSS. So we can extract text without losing quality from it.
- The books in LibGen often have a previous edition (compared to the ones i [redacted]

Caveats:

- Matching algorithm is not perfect as well as the LibGen API (so up to 5% false negatives could be present)
- Sampling from all available titles is not perfect (pseudo random), especially fo [redacted] The problem is that we don't have the full list of titles for either of the publishers, so we need to scrape their web-pages. For that I sampled random

beginning letters and random pages fror[redacted]but the titles are still clustered around certain alphabetic characters

Publisher	Method	Match (%)
[redacted]	Titles&Authors from sampled web-scraping	90% (sample=1000)
[redacted]	Manual check	88% (sample=25)
[redacted]	Titles&Authors from sampled web-scraping	68% (sample=1000)
[redacted]	Manual check	76% (sample=25)

Code:

- Notebook LibGen VS[redacted]VS [redacted]
- [redacted]Web-Scraping
- [redacted]eb-Scraping
- Utils for Web-Schttps://www.internalfb.com[redacted]
- Quick Manual Check: LibGen VS[redacted]VS[redacted]uick manual check

18.04.2023

Motivation: Collect available book titles and authors fro [redacted] nd [redacted]

Observations:

1. We don't have a full list of titles for [redacted] or [redacted] so the following approaches were used:
    a. Web scrape [redacted] and [redacted]
    The problem in this approach is that Springer has 300k book titles and each book has it's own page with details that we need. A lot of requests to be made (possible DDOS)
        i. [redacted] https://link.[redacted]com/books/g/1
        ii. ttps://www[redacted]com/en-us/[redacted]search.html/
    b. Use APIs
    Only [redacted] has APIs for accessing their resources, but it is limited to 100 result per subject. So you first get sample DOI for each category in [redacted] then request details for these DOI. In total you can get 140 books meta (out of 300k) and 2k articles (which we are less interested).
    c. Manual check on their website and randomly checking 25 books
2. One should be careful with doing too many requests to web-resources - I got blocked by LibGen after 1k requests in a few minutes (after I tired multithreading+multiprocessing together).
    a.
3. LibGen API can be missing results (ex. I can find a title manually, but the API doesn't return anything), most likely the API is using a different database. But this is <5% of cases.

Results:

1. Prepared scripts for web-scraping [redacted] and [redacted]

2. Prepared scripts for checking the books in LibGen

Appendix

Links:

- Libgen API: https://pypi.org/project/libgen-api/
- Libgen Search: https://libgen.li
- Sample of documents on fair_cluster: [redacted]atasets/books
- Some description of the project: https://news.ycombinator.com/item?id=21692841
- Libgen Books Metadata: http://libgen.rs/dbdumps/
- [redacted] https://link[redacted]com/books/
- [redacted]ttps://www[redacted]om/en-us[redacted]search.html/

Plan:

1. [in parallel] Find out where to get the dump of the datasets (scitech, fiction and scimag):
    a. Taking metadata from here: http://libgen.rs/dbdumps/
        i. SQL search: It seems that they have the database dumps which I assume are behind the API. It would be much faster to create an SQL database (I assume they use mysql or postgres) which we can setup locally. Then querying is fast.
        ii. Embedding/Elastic search: It might make sense to have some embedding search using fastext embeddings. Encode everything – 100M records with fastext(title), fastext(author), fastext(abstract??). If presented, it would be relatively cheap to search. Then match the concat(ft_title, ft_author, ft_abstract). BoW with wparse char 3-grams should work too.

   - b. Run some high-level stats: share of epub/pdf, share of EN, total count of books, etc...
   - c. Decide on where to store the files: aprox. ~120TB * 30% (english & PDF/EPUB) = ~40TB
   - d. Load the dataset (we probably need filtered data: English and only PDF+EPUB format). Should we load to Meta's Manifold bucket instead of S3?
2. [in parallel] Compare quality of text extraction from LibGen VS [redacted]
   - a. Load samples of pdfs/epubs from the libgen website https://libgen.is/, same samples as from [redacted]
   - b. Check % of [redacted] samples in libgen epub only format
   - c. Parse epub with Marie-Anne Lachaux's html script
   - d. Parse pdf with Lukas's OCR script, record the speed of parsing to further estimate the GPU requirements
   - e. Compare quality VS [redacted] data (original pdfs)
3. [in parallel] Check what books we have in CC (as per Todor Mihaylov's suggestion)
   - a. Check quality/format
   - b. Check intersection with [redacted] titles / LibGen titles (Nikolay Bashlykovto provide code for checking titles using libgen-api)
4. [once data loaded] Filtering & Preprocessing
   - a. Filtering rules
   - b. Run ablations

To Discuss:

- Can we load libgen data using Meta IP ranges? Or should we use some vpn?
  - Redacted - Privilege (to check with Marie-Anne and Guillaume)
- Can we load this data to S3? Or use Meta's Manifold solution? You can load data to RSC from Manifold straightaway and Redacted - Privilege
  - [Mel] Redacted - Privilege Is there any preference to use manifold from a tech perspective? [Todor] No, because we need to process it on AWS/fairspark

- Is there any overlap between the big dump of cc pdfs and libgen pdfs?
  - [Mel] asking so we don't duplicate processing/can prioritize a bit. Maybe easy version is hashing
  - Don't know yet; can try hashing/comparing titles from metadata
- How clean can we get scientific PDFs? Do we still want to buy[redacted]lespite the similarity?
- How long will it take for a first pass of data to be ready?
  - Should we include in v3 or is this too not trending to higher quality models based on our ablations and/or do we feel it is too risky to change our data mix?
  - Should we hold 150B training for this?
  - [Nikolay/Peter] May 17th might for the whole set would be tight; common crawl PDFs seem more doable by then. Just the epub may be possible but need tighter estimates on downloading time (possibly bottlenecked on the p2p network)
  - [Mel] Let's try to batch downloading and processing so we can get some of the data in weeks instead of all of the data in months.
- How much of the datastet is Pdfs? What portion can we use pdf extract for vs need to OCR? how many GPUs is it going to take to OCR the parts of the dataset that can't be pdf extracted for how long (good to know this ASAP)? -> TBD
  - [Peter] 3M books, OCR takes 10 seconds per page/20 mins per book => 1M GPU hours, 3 weeks for 3K GPUs.
  - [Todor] estimate above sounds too high; output might be bad quality
- Still to answer: tighter timeline estimation for first batch of data[redacted]r No, # of GPUs needed when.
  - [Nikolay] estimation for the first batch TBD 29.04 -> run ablation on the first chunk by 12.05
  - [Nikolay] re[redacted]I don't think we need to proceed wit[redacted]t this point:
    - [redacted]verlaps with up to 90% of content in LibGen
    - Quality in LibGen seems to be very high (from a sampled check) for the Sci-tech collection (similar to [redacted] Epub/PDF: 16%/65%
    - LibGen is at least 6 times as large as[redacted] 1.4M books (sci-tech EN books in PDF&Epub) VS 212k (EN books in[redacted]and 32M articles in LibGen VS 3M EN articles in[redacted]

Commented [9]: Remaining things to answer

- [Nikolay] re GPUs needed: with Lukas's estimates on PDF parsing we would need optimally 2k GPUs for OCR parsing to complete in sci-tech in 10 days. And additional 10 days for sci-mag (with less priority). We would need these resources from :
  - fiction: 0
  - sci-tech: 500k GPU*hours
  - sci-mag: 440k GPU*hours (lower priority)
- [Nikolay UPD 28.04] we were able to accelerate the OCR parsing by over 2.5x, so the required GPU*hours would be 2.5x less. We are still analyzing the parsing quality tradeoffs, as this is a smaller model.

Document1

Main document changes and comments

Page 8: Commented [1] **Melanie Kambadur** **10/30/2023 6:58:00 PM**

any rationale of why we're doing this? just better knowledge density? i wonder if it could be useful for long-context?

Page 18: Commented [2] **Melanie Kambadur** **7/31/2023 9:00:00 PM**

Where are we logging results for this? any more details on the experiment?

Page 18: Commented [3] **Nikolay Bashlykov** **8/1/2023 4:02:00 PM**

the main results are below (04.07.2023). this was for the new baseline, but we recently changed it to 4k context length, so this run is not relevant (and was stopped).

I will schedule a new run on the new 4k Dill baseline. But we can also use the previous runs (04.07.2023) - they showed positive signals.

Page 18: Commented [4]

Page 18: Commented [5]

Page 67: Commented [6] **Lukas Blecher** **5/9/2023 1:44:00 PM**

That paints a wrong picture. The speed per GPU is still around 7 books per hour. The number of GPUs is the bottleneck

Page 67: Commented [7] **Nikolay Bashlykov** **5/9/2023 1:50:00 PM**

fair point, can you give a ballpark how much more we need? [redacted]@meta.com was mentioning that we can get 1200 GPUs from the Retina team

Page 67: Commented [8] **Nikolay Bashlykov** **5/16/2023 5:14:00 PM**

As a ball park estimate:
we would need ~2k GPUs on FAIR Cluster for 2 weeks starting from ~mid-next week.
It's actually similar to what we use now for sci-tech, so we might keep the current strategy of just asking people to help run from their accounts.
[redacted]@meta.com, [redacted]meta.com

Page 89: Commented [9] **Melanie Kambadur** **4/24/2023 6:11:00 PM**

Remaining things to answer

Header and footer changes

Text Box changes

Header and footer text box changes

Footnote changes

Endnote changes